Лекція 7. Регресійний аналіз - 1

Данило Тавров

22.03.2023

Вступні зауваги

- Сьогодні ми почнемо розглядати регресійний аналіз, якому присвятимо декілька найближчих лекцій
- Корисними матеріалами є:
  - Фундаментальна книжка *Econometrics* (Bruce Hansen), розділи 2–4, 7 (викладено на диску в загальному каталозі з літературою)
    - Це справді дуже детальна книжка, тому з цих розділів потрібно не все
    - Зверніть увагу тільки на ті моменти, які ми охоплюємо в лекції
  - Книжка *Introduction to Econometrics* (James H. Stock, Mark W. Watson), розділи 4, 6 (викладено на диску в загальному каталозі з літературою)
- Матеріал цієї лекції частково базується на конспекті лекцій із дисципліни ECON 141 *Econometrics: Math Intensive* (University of California, Berkeley) авторства Віри Семенової та Данила Таврова

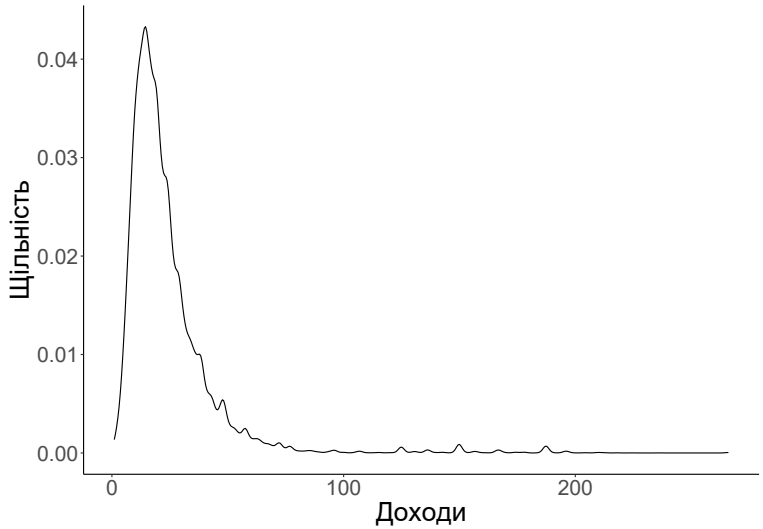
План лекції

Мотивація

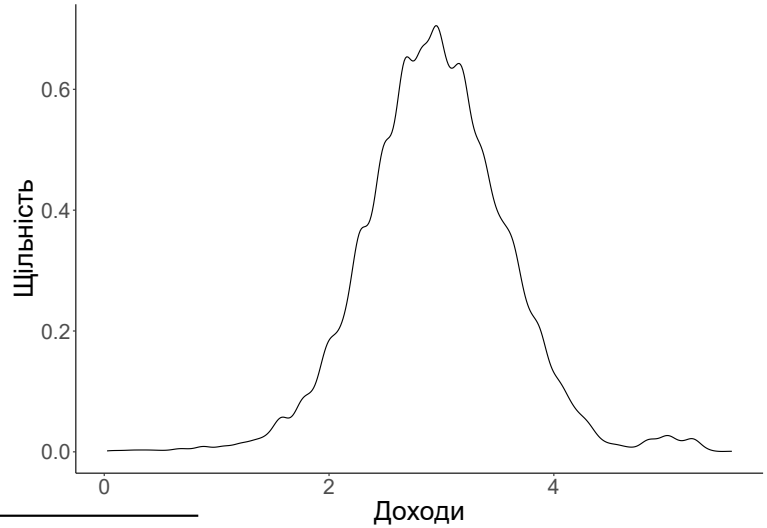
- Проводячи аналіз даних, ідеальний результат, який ми волiємо дістати — це встановити **причиново-наслідковий** зв'язок між різними змінними
- Нас цікавить питання: якщо змінити одну змінну на декілька одиниць, то як сильно зміниться інша змінна?
  - Чи впливає кількість учнів у класі на показники в навчанні
  - Чи впливає рівень видатків на поліцію на рівень злочинности
  - Чи впливає рівень освіти на зарплату
  - Чи впливає метод лікування на здоров'я
  - Тощо
- Ці та схожі питання є питаннями *ceteris paribus* (із латини — *за інших рівних умов*)
- Іншими словами, ми хочемо проаналізувати вплив однієї змінної, вважаючи, що інші змінні залишаються фіксованими
  - Звісно, проаналізувати **конкретний** вплив просто неможливо, бо в нас немає даних про одні й ті самі одиниці спостереження *з* і *без* деякого фактора
  - Наприклад, у працівника є конкретний рівень освіти — ми не спостерігаємо, *що могло б бути*, якби в нього був інший рівень
  - Тому ми можемо давати відповідь тільки в певному «середньому» сенсі
- Регресійний аналіз — надзвичайно поширений спосіб (спроби) дати відповідь на такі питання

Сподівання та умовне сподівання (1)

- Розгляньмо дані про зарплати працівників США (березень 2009 р.)[1]
- Можемо оцінити щільність погодинних доходів



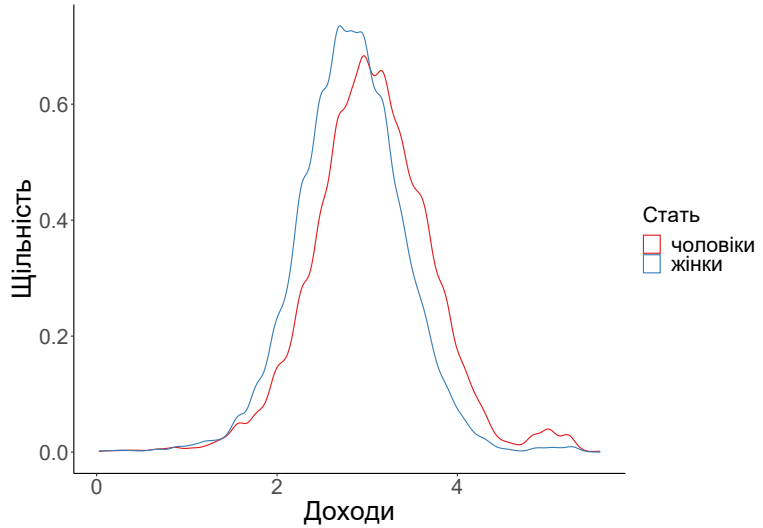
- Як бачимо, розподіл є суттєво скошеним, тому доцільно перейти до логаритмів



[1] Дані Current Population Survey (CPS) про 57 000 домогосподарств з сайту Брюса Хансена

Сподівання та умовне сподівання (2)

- Можемо розглянути сподівання логаритмів зарплат як показник певної «типової» зарплати
- Так, ми маємо, що $\mathbb{E}\left[\ln \text{wage}\right] \approx 2.956$[2]
- Проте цієї інформації недостатньо, щоб зрозуміти справжню картину
  - Хотілося б знати, як залежить сподівання від різних характеристик працівників
- Наприклад, можна проаналізувати розподіл зарплат для жінок і для чоловіків



- Розподілі доволі подібні, тільки для чоловіків він зсунутий управо
- Можна обчислити відповідні **умовні** сподівання:

```
wages_summary <- wages %>% group_by(female) %>% summarise(mean_wage = mean(log_hourly_wage))
```

  - $\mathbb{E}\left[\ln \text{wage} \mid \text{female} = 0\right] \approx 3.058$
  - $\mathbb{E}\left[\ln \text{wage} \mid \text{female} = 1\right] \approx 2.818$

[2] До речи, зверніть увагу, що $e^{\mathbb{E}[\ln \text{wage}]}$ фактично є середнім геометричним зарплат

Сподівання та умовне сподівання (3)

- Понад те, ми вже володіємо інструментарієм, який може визначити, чи є ця різниця статистично значущою

```
t.test(log_hourly_wage ~ female, data = wages)
```

```
##
##  Welch Two Sample t-test
##
## data:  log_hourly_wage by female
## t = 43.024, df = 49107, p-value < 2.2e-16
## alternative hypothesis: true difference in means is not equal to 0
## 95 percent confidence interval:
##  0.2285041 0.2503176
## sample estimates:
## mean in group 0 mean in group 1
##        3.057743        2.818332
```

- Як можна бачити, різниця надзвичайно статистично значуща
  - Зверніть увагу на 95% довірчий інтервал: він доволі вузький, бо $n = 50628$
- Оскільки це різниця в **логаритмах**, потрібно коректно її інтерпретувати
  - **Грубо** кажучи, середні зарплати чоловіків у $e^{3.06-2.82} \approx 1.27$ разів вищі
  - Ми до цього повернемося в наступній лекції
- Потрібно усвідомлювати, що на цьому етапі ми **не можемо** нічого сказати про **причиново-наслідковий зв'язок** між статтю та зарплатою
  - Поки що це просто статистичне спостереження про різницю між двома розподілами

Сподівання та умовне сподівання (4)

- На цьому наші знання з попередніх лекцій завершуються
- Але цілком очевидно, що також цікавим є питання про умовні сподівання за умов **декількох** змінних
- Наприклад, можна додати, чи є працівник латиноамериканцем

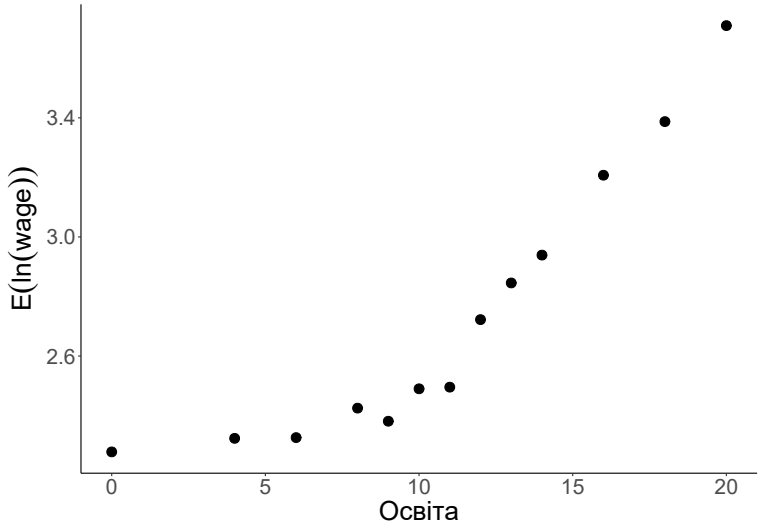
```
wages_summary_hisp <- wages %>% group_by(female, hisp) %>%
  summarise(mean_wage = mean(log_hourly_wage))
wages_summary_hisp
```

```
## # A tibble: 4 x 3
## # Groups:   female [2]
##   female  hisp mean_wage
##    <dbl> <dbl>     <dbl>
## 1      0     0      3.12
## 2      0     1      2.72
## 3      1     0      2.86
## 4      1     1      2.58
```

- Наприклад, $\mathbb{E}\left[\ln \text{wage} \mid \text{female} = 0, \text{hisp} = 0\right] \approx 3.12$
  - Як можна бачити, латиноамериканці заробляють іще менше
  - Можна порівняти ці середні попарно між групами і виконати t-тести
  - Хотілося б мати можливість робити висновки, враховуючи всю інформацію
- Таких сподівань можна побудувати дуже багато
  - Вони проливають світло на особливості зарплат для різних категорій працівників
  - Різниці між (середніми) зарплатами можуть свідчити про дискримінацію
  - Аналіз впливу різних характеристик на зарплату може пояснювати, що важливе для ринку праці тощо
- При цьому варто розуміти, що просто умовні сподівання самі по собі є всього лише дескриптивними статистиками, і вони нічого не пояснюють

Сподівання та умовне сподівання (5)

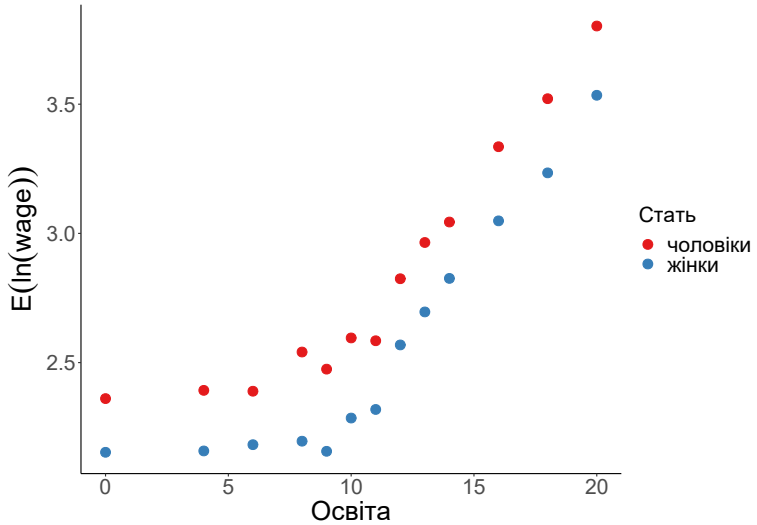
- Повернімося до нашого початкового питання
- Ми хотіли з'ясувати, чи є вплив освіти на зарплату
- У цьому датасеті під *освітою* розуміють кількість років, проведених у закладах освіти після дитсадка
  - Так, зокрема, для середньої школи маємо 12
  - Для бакалавра — 16 тощо
- Можна збудувати відповідні графіки умовних сподівань



- Спостерігаємо зростання зарплати залежно від рівня освіти
  - До того ж швидкість зростання суттєво вища після школи

Сподівання та умовне сподівання (6)

- Також можна додати інформацію про стать і побудувати графік умовних сподівань окремо для чоловіків і жінок



- Можна бачити, що в жінок (середні) зарплати нижчі для всіх рівнів освіти
- Звісно, можна додати інші характеристики — расу, сімейний стан, місце роботи тощо — і дістати ще детальнішу картину

Функція умовного сподівання

- Це, власне, приводить нас до поняття **функції умовного сподівання** (conditional expectation function, CEF):

$$\mathbb{E}\left[Y \mid \mathbf{X} = \mathbf{x}\right] = \mathbb{E}\left[Y \mid X_1 = x_1, \dots, X_k = x_k\right] \equiv m(x_1, \dots, x_k) \tag{1.1}$$

- У цьому контексті Y називають **залежною змінною** (dependent variable), X_i — **незалежними змінними** (independent variables)

Функція умовного сподівання як найліпший предиктор (1)

- Ми можемо довести, що CEF є **найліпшим предиктором** (best predictor) Y
  - У тому сенсі, що воно мінімізує значення середньоквадратичної похибки:

$$m(\mathbf{X}) = \mathbb{E}\left[Y \mid \mathbf{X}\right] = \arg\min_{g} \mathbb{E}\left[(Y - g(\mathbf{X}))^2\right] \tag{1.2}$$

  - Для цього потрібно, щоб Y мала скінченний момент другого порядку
- Справді,

$$\begin{aligned}\mathbb{E}\left[(Y - g(\mathbf{X}))^2\right] &= \mathbb{E}\left[(Y - m(\mathbf{X}) + m(\mathbf{X}) - g(\mathbf{X}))^2\right] \\ &= \mathbb{E}\left[(Y - m(\mathbf{X}))^2\right] + 2\mathbb{E}\left[(Y - m(\mathbf{X}))(m(\mathbf{X}) - g(\mathbf{X}))\right] \\ &+ \mathbb{E}\left[(m(\mathbf{X}) - g(\mathbf{X}))^2\right]\end{aligned}$$

Функція умовного сподівання як найліпший предиктор (2)

- За законом ітерованих сподівань,

$$\mathbb{E}\left[(Y - m(\mathbf{X}))(m(\mathbf{X}) - g(\mathbf{X}))\right] = \mathbb{E}\left[\mathbb{E}\left[(Y - m(\mathbf{X}))(m(\mathbf{X}) - g(\mathbf{X}))\right] \mid \mathbf{X}\right]$$

- Можна помітити, що цей вираз дорівнює

$$(m(\mathbf{X}) - g(\mathbf{X})) \cdot \mathbb{E}\left[Y - m(\mathbf{X}) \mid \mathbf{X}\right] = (m(\mathbf{X}) - g(\mathbf{X})) \cdot (\mathbb{E}\left[Y \mid \mathbf{X}\right] - m(\mathbf{X})) = 0$$

- Вираз $\mathbb{E}\left[(Y - m(\mathbf{X}))^2\right]$ взагалі не залежить від $g(\mathbf{X})$
- Відтак $\mathbb{E}\left[(Y - g(\mathbf{X}))^2\right]$ досягатиме мінімуму, якщо $\mathbb{E}\left[(m(\mathbf{X}) - g(\mathbf{X}))^2\right]$ буде найменшим
- Тобто якщо $g(\mathbf{X}) = m(\mathbf{X})$

Похибка CEF (1)

- Цілком очевидно, що CEF мінімізує *середньо*квадратичну похибку
- Для кожного *конкретного* спостереження в загальному випадку матиме місце **похибка** (error):

$$e = Y - m(\mathbf{X})$$

- Корисна властивість похибки CEF випливає з властивостей умовних сподівань:

$$\mathbb{E}\left[e \mid \mathbf{X}\right] = \mathbb{E}\left[Y - m(\mathbf{X}) \mid \mathbf{X}\right] = 0$$

- Іншими словами, можемо записати таку **регресійну** модель:

$$\begin{aligned} Y &= m(\mathbf{X}) + e \\ \mathbb{E}\left[e \mid \mathbf{X}\right] &= 0 \end{aligned} \tag{1.3}$$

- Варто зазначити, що ніде не вимагається, щоб $e \perp\!\!\!\perp \mathbf{X}$
  - Потрібно тільки, щоб умовне сподівання дорівнювало 0
  - Тобто для кожного значення $\mathbf{X}$ в середньому відхилення від CEF нульові

Похибка CEF (2)

- Використавши закон ітерованих сподівань, маємо

$$\mathbb{E}[e] = \mathbb{E}[\mathbb{E}[e \mid \mathbf{X}]] = 0$$

- Отже не тільки умовне, але й *без*умовне сподівання e дорівнює 0
- Аналогічно можна довести[3] таку корисну властивість:

$$\mathbb{E}[h(\mathbf{X})e] = 0$$

- Відтак $\text{Cov}(h(\mathbf{X}), e) = \mathbb{E}[h(\mathbf{X})e] - \mathbb{E}[h(\mathbf{X})]\,\mathbb{E}[e] = 0 - 0 = 0$
  - Тобто похибка некорельована з **будь-якою** функцією від $\mathbf{X}$
  - При цьому це **зовсім не означає**, що похибка **незалежна** від $\mathbf{X}$
- За принципом контрапозиції відразу випливає, що якщо $\mathbb{E}[h(\mathbf{X})e] \neq 0$, то обов'язково $\mathbb{E}[e \mid \mathbf{X}] \neq 0$
- Тобто, зокрема, **якщо похибка корелює з** X_i, то умова про нульове сподівання **не виконується**
- Також можна показати, що якщо $\mathbb{E}[|Y|^r] < \infty$, то й $\mathbb{E}[|e|^r] < \infty$
  - Тобто якщо існує деякий момент для Y, то він існує і для e

[3] Доведіть!

Безумовне сподівання як найліпший предиктор

- Суто формально регресійна модель може взагалі не містити $\mathbf{X}$:

$$Y = \mu + e$$
$$\mathbb{E}[e] = 0$$

- У цьому випадку CEF є просто $\mathbb{E}[Y] = \mathbb{E}[\mu + e] = \mu$
- Відтак можемо стверджувати, що **безумовне** сподівання є найліпшим предиктором Y у розумінні мінімізації $\mathbb{E}\left[(Y - b)^2\right]$ для всіх $b \in \mathbb{R}$

Причиново-наслідкова інтепретація (1)

- Нехай ми маємо модель $Y = m(\mathbf{X}) + e$, $\mathbb{E}[e \mid \mathbf{X}] = 0$
  - Це друге рівняння стверджує, що m є саме CEF для Y, а не просто якась функція
- Якщо ми віримо, що ця модель є істинною, то тоді ніщо нам не заважає провести аналіз *ceteris paribus*
  - Звісно, може бути таке, що ми думаємо, що X впливає на Y, хоча насправді Y впливає на X
  - У багатьох випадках очевидно, про що мова (зарплата не може впливати на стать тощо)
  - Але ми до цього також повернемося далі в нашому курсі
- Наприклад, можемо мати модель $Y = m(X_1, X_2) + e$, $\mathbb{E}[e \mid X_1, X_2] = 0$
  - Тут, скажімо, Y — зарплата, X_1 — освіта, X_2 — стать
- Якщо ми вважаємо, що ця модель істинна, то з урахуванням вищенаведених міркувань ми вважаємо, що **жодна інша змінна не корелює** ні з X_1, ні з X_2
  - Наскільки це реалістично?!
- Менше з тим, якщо ми вважаємо, що це так, то ми можемо говорити про вплив **винятково** освіти
  - Наприклад, можемо казати, що зміна освіти з 12 до 13 років має наслідком (у середньому) певне збільшення зарплати
  - Кажемо, що стать у цьому випадку є **контрольованою** (controlled for)
- Тоді $\frac{\partial \mathbb{E}[Y|X_1,X_2]}{\partial X_1} = \frac{\partial m}{\partial X_1}$
  - Або, якщо X_1 дискретна, то $\frac{\partial \mathbb{E}[Y|X_1,X_2]}{\partial X_1} = m(X_1 = a+1, X_2 = b) - m(X_1 = a, X_2 = b)$ тощо

Причиново-наслідкова інтепретація (2)

- У чому тоді проблема?
- Ми **не знаємо**, чому дорівнює m!
- Щонайменше ми не знаємо функціональної форми
  - Тобто незрозуміло, як рахувати відповідну похідну
- На практиці ми вимушені **припускати**, що m належить певному класу
- Найчастіше таким класом є клас **лінійних** функцій
  - Якщо лінійна апроксимація доволі непогана, то в цьому проблеми великої немає
  - Зрештою, гладку функцію можна розвинути в ряд Тейлора з довільною точністю
  - Проблеми виникатимуть із **оцінювання** таких моделей
  - Але спочатку потрібно зрозуміти, що це за моделі

План лекції

Лінійна CEF

- Позначмо $\mathbf{X} = (1, X_1, \dots, X_k)^\top$
- Тоді **лінійна** CEF дорівнює

$$\mathbb{E}[Y \mid \mathbf{X}] \equiv m(\mathbf{X}) = \beta_0 + \sum_{i=1}^{k} \beta_k X_k = \mathbf{X}^\top \beta \,, \quad \beta = (\beta_0, \beta_1, \dots, \beta_k)^\top \tag{2.1}$$

- Відповідно, **лінійна регресійна модель** дорівнює

$$\begin{aligned} Y &= \mathbf{X}^\top \beta + e \\ \mathbb{E}[e \mid \mathbf{X}] &= 0 \end{aligned} \tag{2.2}$$

Інтерпретація коефіцієнтів лінійної моделі

- Нехай наша регресійна модель справді має причиново-наслідкову інтерпретацію
  - Тобто справді $\mathbb{E}\left[e \mid \mathbf{X}\right] = 0$
- Тоді вплив кожної незалежної змінної X_i, *ceteris paribus*, просто дорівнює значенню відповідного коефіцієнта β_i
- «Лінійність» модели означає, що вона лінійна в коефіцієнтах, а не в регресорах
- Наприклад, може бути модель $Y = \beta_0 + \beta_1 X_1 + \beta_2 X_2 + \beta_3 X_2^2 + e$
  - Суто формально вектор регресорів є $\mathbf{X} = (1, X_1, X_2, X_2^2)^\top$
  - І сама модель залишається лінійною: $Y = \mathbf{X}^\top \beta$
- Якщо в моделі присутні нелінійні функції від X_i, то тоді відповідні похідні $\frac{\partial m(\mathbf{X})}{\partial X_i}$ потрібно рахувати окремо
  - Коефіцієнти β_i вже не будуть мати такої прямої інтерпретації
  - Наприклад, $\frac{\partial \mathbb{E}[Y|\mathbf{X}]}{\partial X_2} = \beta_2 + 2X_2\beta_3$
  - Ми до цього повернемося в наступній лекції

Лінійна модель як апроксимація CEF (1)

- Уявімо собі, що в нас тільки одна бінарна змінна X, яка набуває всього двох значень — 1 і 0
  - Наприклад, стать
  - Такі змінні називають **індикаторними** (indicator) або **фіктивними** (dummy)
- Тоді, вочевидь, ми маємо тільки два значення, $\mathbb{E}[Y \mid X = 0]$ і $\mathbb{E}[Y \mid X = 1]$
- Відповідно, CEF **не може не бути лінійною**:

$$\mathbb{E}[Y \mid X] = \beta_0 + \beta_1 X$$

- Тут $\beta_0 = \mathbb{E}[Y \mid X = 0]$, $\beta_1 = \mathbb{E}[Y \mid X = 1] - \mathbb{E}[Y \mid X = 0]$
- Між іншим, принципово, що саме закодовано як $X = 1$
  - Якщо $X = 1$ відповідає чоловікам, то тоді β_0 — це умовне сподівання для жінок
  - Якщо ж $X = 1$ відповідає жінкам, то все навпаки

Лінійна модель як апроксимація CEF (2)

- Нехай тепер маємо дві бінарні змінні, X_1 та X_2
  - Наприклад, стать і одружений/неодружений
- Тоді може бути всього 4 можливі значення для $\mathbb{E}[Y \mid X_1, X_2]$
- І знову CEF **не може не бути лінійною**:

$$\mathbb{E}[Y \mid X_1, X_2] = \beta_0 + \beta_1 X_1 + \beta_2 X_2 + \beta_3 X_1 X_2$$

  - $\beta_0 = \mathbb{E}[Y \mid X_1 = 0, X_2 = 0]$
  - $\beta_1 = \mathbb{E}[Y \mid X_1 = 1, X_2 = 0] - \mathbb{E}[Y \mid X_1 = 0, X_2 = 0]$
  - $\beta_2 = \mathbb{E}[Y \mid X_1 = 0, X_2 = 1] - \mathbb{E}[Y \mid X_1 = 0, X_2 = 0]$
  - $\beta_3 = \mathbb{E}[Y \mid X_1 = 1, X_2 = 1] - \mathbb{E}[Y \mid X_1 = 0, X_2 = 1] - \mathbb{E}[Y \mid X_1 = 1, X_2 = 0] + \mathbb{E}[Y \mid X_1 = 0, X_2 = 0]$
- Змінну $X_1 X_2$ називають **фактором взаємодії** (interaction term) між X_1 та X_2
- Вочевидь, продовжуючи ці міркування, маємо, що для k бінарних змінних CEF повинна бути лінійною
  - Вона міститиме 2^k регресорів
  - Серед них $X_1, \dots, X_k$ та всі можливі взаємодії

Лінійна модель як апроксимація CEF (3)

- Нехай деяка змінна є категорійною, тобто $X \in \{x_1, \dots, x_p\}$
  - Наприклад, регіон проживання
- Тоді її можна перетворити в $p - 1$ бінарні змінні
  - Кожна нова змінна V_i дорівнює 1 або 0 залежно від того, чи дорівнює X значенню x_i
  - Зверніть увагу, що змінної V_p нам непотрібно
  - Адже якщо $X = x_p$, то це означає, що $V_1 = \dots = V_{p-1} = 0$
- Тоді CEF також **обов'язково** буде лінійна:

$$\mathbb{E}[Y \mid X] = \beta_0 + \sum_{i=1}^{p-1} \beta_i V_i$$

- Зрозуміло, що якщо є декілька змінних, як бінарних, так і категорійних, то можна утворити відповідну лінійну CEF
  - Такі моделі називають **насиченими** (saturated)
  - Число коефіцієнтів дорівнює числу можливих значень, які набувають наші змінні

Лінійна модель як апроксимація CEF (4)

- Але в загальному випадку цей підхід особливо далеко не заведе
  - Мало яка модель містить у собі тільки дискретні змінні
  - Якщо змінних (і значень, яких вони набувають) доволі багато, то насичена модель може бути дуже великою
- Відтак лінійні моделі найчастіше є **апроксимаціями** справжньої (невідомої нам) CEF
- Зрозуміло, що нас цікавить не будь-яка довільна лінійна модель з незрозумілими коефіцієнтами
- Нас цікавить лінійна модель $Y = \mathbf{X}^\top \beta$, яка є **найліпшим (лінійним) предиктором** Y
  - У тому сенсі, що вона мінімізує MSE $\mathbb{E}\left[\left(Y - X^\top \mathbf{b}\right)^2\right]$ за всіма $\mathbf{b} \in \mathbb{R}^k$
- Якщо так задуматися, то фактично $\mathbf{X}^\top \beta$ є **проєкцією** Y на лінійний простір, утворений із $\mathbf{X}$
  - Коефіцієнти β називають **коефіцієнтами проєкції** (projection coefficients)

Лінійна модель як апроксимація CEF (5)

- Для виведення цих коефіцієнтів потрібно здійснити такі **припущення**:
  - $\mathbb{E}\left[Y^2\right] < \infty$
  - $\mathbb{E}\left[\|\mathbf{X}\|^2\right] < \infty$[4]
    - Ці дві умови потрібні, щоб гарантувати існування всіх сподівань та коваріацій
    - Зокрема, $\|\mathbb{E}\left[\mathbf{X}Y\right]\| \leq \mathbb{E}\left[\|\mathbf{X}Y\|\right]$ за нерівністю Єнсена
    - І тоді $\mathbb{E}\left[\|\mathbf{X}Y\|\right] \leq \sqrt{\mathbb{E}\left[\|\mathbf{X}\|^2\right] \cdot \mathbb{E}\left[|Y|^2\right]} < \infty$ за нерівністю Коші-Буняковського
    - Відтак $\|\mathbb{E}\left[\mathbf{X}Y\right]\| < \infty$, що гарантує скінченність **кожного** сподівання $\mathbb{E}\left[X_i Y\right]$, $i = 1, \dots, k$
    - Аналогічно можна показати для $\mathbb{E}\left[\mathbf{X}\mathbf{X}^\top\right]$, використавши норму матриці як корінь із суми квадратів усіх її елементів
  - Матриця $\mathbf{Q}_{\mathbf{XX}} \equiv \mathbb{E}\left[\mathbf{X}\mathbf{X}^\top\right]$ є додатно визначеною
- Тоді

$$\text{MSE} = \mathbb{E}\left[Y^2\right] - 2\beta^\top \mathbb{E}\left[\mathbf{X}Y\right] + \beta^\top \mathbb{E}\left[\mathbf{X}\mathbf{X}^\top\right]\beta \equiv \mathbb{E}\left[Y^2\right] - 2\beta^\top \mathbf{Q}_{\mathbf{X}Y} + \beta^\top \mathbf{Q}_{\mathbf{XX}}\beta$$

- Умова першого порядку дорівнює

$$0 = \frac{\partial \text{MSE}}{\partial \beta} = -2\mathbf{Q}_{\mathbf{X}Y} + 2\mathbf{Q}_{\mathbf{XX}}\beta$$

- Звідси

$$\beta = \mathbf{Q}_{\mathbf{XX}}^{-1}\mathbf{Q}_{\mathbf{X}Y} = \left(\mathbb{E}\left[\mathbf{X}\mathbf{X}^\top\right]\right)^{-1}\mathbb{E}\left[\mathbf{X}Y\right] \tag{2.3}$$

  - ~~Умову другого порядку пере~~вірте самостійно

[4] Тут $\|\cdot\|$ є евклідовою нормою вектора: $\|\mathbf{X}\| = \sqrt{\mathbf{X}^\top\mathbf{X}}$

Лінійна модель як апроксимація CEF (6)

- Тепер зрозуміло, навіщо додатна визначеність матриці $\mathbf{Q_{XX}}$
  - Ця матриця завжди невід'ємно визначена: $\mathbf{a}^\top \mathbf{Q_{XX}} \mathbf{a} = \mathbb{E}\left[\left(\mathbf{a}^\top \mathbf{X}\right)^2\right] \geq 0$
  - Вимога строгої визначености виключає можливість нульового розв'язку рівняння
- Отже якщо виконуються вказані вище припущення, β буде єдиним
  - Кажуть, що параметр β **ідентифікований** (identified)
  - Тобто його можна в єдиний спосіб обчислити
- Можливою **оцінкою** $\widehat{\beta}$ може бути plug-in оцінка, де ми замінюємо всі сподівання на вибіркові аналоги
  - Нехай маємо вибірку $(Y_i, X_{1,i}, \dots, X_{k,i}), i = 1, \dots, n$, $\mathbf{X}_i = (1, X_{1,i}, \dots, X_{k,i})^\top$
  - Тоді plug-in оцінкою цих коефіцієнтів буде

$$\widehat{\beta} = \left(\frac{1}{n}\sum_{i=1}^{n} \mathbf{X}_i \mathbf{X}_i^\top\right)^{-1} \left(\frac{1}{n}\sum_{i=1}^{n} \mathbf{X}_i Y_i\right)$$

  - Ми про це говоритимемо далі

Лінійна проєкція

- Відтак маємо найліпший лінійний предиктор
  $\mathcal{P}(Y \mid \mathbf{X}) = \mathbf{X}^\top \beta = \mathbf{X}^\top \left(\mathbb{E}\left[\mathbf{X}\mathbf{X}^\top\right]\right)^{-1} \mathbb{E}\left[\mathbf{X}Y\right]$
  - Його називають **лінійною проєкцією** (linear projection)
- Тоді **похибкою проєкції** (projection error) є

$$e = Y - \mathbf{X}^\top \beta$$

  - Рівняння $Y = \mathbf{X}^\top \beta + e$ часто називають **регресією** (regression) Y на $\mathbf{X}$
- Нескладно показати

$$\mathbb{E}\left[\mathbf{X}e\right] = \mathbb{E}\left[\mathbf{X}\left(Y - \mathbf{X}^\top \beta\right)\right] = \mathbf{Q}_{\mathbf{X}Y} - \mathbf{Q}_{\mathbf{X}\mathbf{X}}\mathbf{Q}_{\mathbf{X}\mathbf{X}}^{-1}\mathbf{Q}_{\mathbf{X}Y} = 0$$

- Оскільки $\mathbf{X} = (1, X_1, \dots, X_k)^\top$, то фактично це означає, що $\mathbb{E}\left[e\right] = 0$, $\mathbb{E}\left[X_1 e\right] = 0$, ..., $\mathbb{E}\left[X_k e\right] = 0$
  - Зокрема похибка проєкції в середньому дорівнює 0
- Також можна бачити, що $\text{Cov}\left(X_j, e\right) = \mathbb{E}\left[X_j e\right] - \mathbb{E}\left[X_j\right]\mathbb{E}\left[e\right] = 0, j = 1, \dots, k$
  - Отже похибка проєкції некорельована з усіма регресорами
- Тобто **якщо похибка** e в рівнянні $Y = \mathbf{X}^\top \beta + e$ **корельована** з $\mathbf{X}$, то це **не є лінійна проєкція!**

Альтернативний запис коефіцієнтів проєкції (1)

- Візьмімо в нашій моделі сподівання:

$$\mathbb{E}[Y] = \beta_0 + \sum_{i=1}^{k} \beta_i \mathbb{E}[X_i] + \mathbb{E}[e] = \beta_0 + \sum_{i=1}^{k} \beta_i \mathbb{E}[X_i]$$

- Відтак $\beta_0 = \mathbb{E}[Y] - \sum_{i=1}^{n} \beta_i \mathbb{E}[X_i]$
- Тоді $Y - \mathbb{E}[Y] = \sum_{i=1}^{k} \beta_i (X_i - \mathbb{E}[X_i]) + e$
- Похибка e некорельована з будь-яким X_i, а тому $\mathbb{E}[(X_i - \mathbb{E}[X_i])\, e] = 0$
- Отже можна застосувати формулу лінійної проєкції

Альтернативний запис коефіцієнтів проєкції (2)

- Матимемо

$$\begin{pmatrix}\beta_1\\ \vdots\\ \beta_k\end{pmatrix} = \left(\mathbb{E}\left[\begin{pmatrix}X_1-\mathbb{E}[X_1]\\ \vdots\\ X_k-\mathbb{E}[X_k]\end{pmatrix}(X_1-\mathbb{E}[X_1],\dots,X_k-\mathbb{E}[X_k])\right]\right)^{-1}$$
$$\times\,\mathbb{E}\left[\begin{pmatrix}X_1-\mathbb{E}[X_1]\\ \vdots\\ X_k-\mathbb{E}[X_k]\end{pmatrix}(Y-\mathbb{E}[Y])\right]$$
$$= \mathbf{Cov}\left(\begin{pmatrix}X_1\\ \vdots\\ X_k\end{pmatrix}\right)^{-1}\mathrm{Cov}\left(\begin{pmatrix}X_1\\ \vdots\\ X_k\end{pmatrix},Y\right)$$

  - Тобто коефіцієнти лінійної проєкції залежать тільки від відповідних коваріацій
  - Зокрема, наприклад, якщо маємо всього один регресор, $Y_i = \beta_0 + \beta_1 X_i + u_i$, то

$$\beta_1 = \frac{\mathrm{Cov}(X_i, Y_i)}{\mathrm{Var}(X_i)}$$

Різниця між CEF та проєкціями (1)

- Перед тим, як перейти до оцінювання коефіцієнтів проєкції, потрібно **дуже чітко** усвідомити різницю між функцією умовного сподівання та лінійною проєкцією
- Отже ми хочемо встановити, яка є залежність між Y та $\mathbf{X}$
  - В ідеалі ми хочемо довести причиново-наслідковий зв'язок між ними
  - Нас цікавить, як змінюватиметься Y, якщо змінити один із X_i *ceteris paribus*, тобто тримаючи всі інші регресори фіксованими
- Функція умовного сподівання CEF $m(\mathbf{X}) = \mathbb{E}[Y \mid \mathbf{X}]$ мінімізує середньоквадратичну похибку між Y та цією функцією
- Відповідна модель має вигляд $Y = m(\mathbf{X}) + e$, $\mathbb{E}[e \mid \mathbf{X}] = 0$
  - Рівність нулю принципова: якщо її немає, то рівняння може й не бути рівнянням CEF
- На практиці ми не можемо встановити, чому дорівнює m
- Тому як апроксимацію ми використовуємо лінійну CEF $Y = \mathbf{X}^\top \beta + \tilde{e}$, $\mathbb{E}[\tilde{e} \mid \mathbf{X}] \approx 0$
  - Звісно, CEF може бути лінійною з самого початку, наприклад, коли всі регресори дискретні, і модель насичена
  - Коефіцієнти β мають інтерпретацію сили зв'язку між регресорами та Y
  - У загальному випадку CEF не є лінійною, і тоді формально ми маємо $Y = \mathbf{X}^\top \beta + \big(m(\mathbf{X}) - \mathbf{X}^\top \beta + e\big)$
  - Тоді $\mathbb{E}[Y \mid \mathbf{X}] = \mathbf{X}^\top \beta + \big(m(\mathbf{X}) - \mathbf{X}^\top \beta\big) \approx \mathbf{X}^\top \beta$, якщо $m(\mathbf{X}) \approx \mathbf{X}^\top \beta$
  - У цьому випадку коефіцієнти β варто сприймати як певну **апроксимацію** такого зв'язку

Різниця між CEF та проєкціями (2)

- **Обчислити** β ми можемо як відповідні коефіцієнти лінійної проєкції $Y = \mathbf{X}^\top \beta + u$
- Але для лінійної проєкції справедливо $\mathbb{E}[\mathbf{X}u] = 0$ суто за побудовою
- Відтак якщо ця умова на практиці не виконується, то **коефіцієнти будуть неправильні!**
- Скажімо, нас може цікавити вплив стати X_1, раси X_2 та освіти X_3 на зарплату Y
- Справжня модель може бути

$$Y = h(X_1, X_2, X_3) + e\,, \qquad \mathbb{E}[e \mid X_1, X_2, X_3] = 0$$

- Ми можемо записати її лінійну апроксимацію:

$$Y = \beta_0 + \beta_1 X_1 + \beta_2 X_2 + \beta_3 X_3 + \tilde{e}\,, \qquad \mathbb{E}[\tilde{e} \mid X_1, X_2, X_3] \approx 0$$

- Ми можемо оцінити її коефіцієнти β як коефіцієнти лінійної проєкції
  - **Якщо** $\tilde{e}$ некорельована з X_1, X_2, X_3, то відповідні коефіцієнти $\beta_1, \beta_2, \beta_3$ будуть спроможними (далі розглянемо це детально)
- Ми можемо також записати **альтернативну** апроксимацію:

$$Y = \gamma_0 + \gamma_1 X_1 + \gamma_2 X_2 + \gamma_3 X_1 X_2 + \gamma_4 X_3 + \check{e}\,, \qquad \mathbb{E}[\check{e} \mid X_1, X_2, X_3] \approx 0$$

  - Ми можемо також оцінити γ як коефіцієнти, але вже іншої лінійної проєкції
  - **Якщо** $\tilde{e}$ некорельована з X_1, X_2, X_1X_2, X_3, то відповідні коефіцієнти $\gamma_1, \gamma_2, \gamma_3, \gamma_4$ будуть спроможними
- **Допоки** виконується умова, що похибка має умовне нульове сподівання, мова тільки про якість апроксимації

Ілюстративні приклади (1)

- Повернімося до нашого прикладу з зарплатами

```
wages_summary_hisp
```

```
## # A tibble: 4 x 3
## # Groups:   female [2]
##   female  hisp mean_wage
##    <dbl> <dbl>     <dbl>
## 1      0     0      3.12
## 2      0     1      2.72
## 3      1     0      2.86
## 4      1     1      2.58
```

- Розгляньмо питання оцінювання CEF для стати X_1 та раси X_2
- Як ми вже знаємо, така CEF буде справді лінійною:

$$\mathbb{E}\left[Y \mid X_1, X_2\right] = 3.12 - 0.263X_1 - 0.398X_2 + 0.123X_1X_2$$

- Зокрема, скажімо, -0.26 — це різниця середніх (логаритмів) зарплат між чоловіками-латиноамериканцями ($X_1 = 0, X_2 = 1$) та білими чоловіками ($X_1 = X_2 = 0$)
- **Уявімо на хвилинку**, що більше в нашому світі немає жодних інших змінних, тільки стать і вік
- Тоді можна записати модель $Y = \beta_0 + \beta_1 X_1 + \beta_2 X_2 + \beta_3 X_1 X_2 + e$, $\mathbb{E}\left[e \mid X_1, X_2\right] = 0$
  - Звідси випливає $\mathbb{E}\left[e \cdot (X_1, X_2)^\top\right] = 0$, тобто похибка некорельована з регресорами
  - Відтак можна оцінити цю модель за допомогою формули лінійної проєкції

Ілюстративні приклади (2)

- Можемо розглянути тепер модель, у якій не враховано терм взаємодії $X_1 X_2$:

$$\mathcal{P}(Y \mid X_1, X_2) = 3.112 - 0.245 X_1 - 0.349 X_2$$

- Це вже **не буде** CEF (CEF було пораховано на попередньому слайді)
  - Це всього лише лінійна апроксимація
  - Фактично маємо $Y = \beta_0 + \beta_1 X_1 + \beta_2 X_2 + (\beta_3 X_1 X_2 + e)$
  - Її можна порахувати як лінійну проєкцію Y на X_1, X_2
  - Але оскільки X_1 і X_2 корельовані з $X_1 X_2$, ці коефіцієнти будуть **некоректні**
- Справді, як можна бачити, коефіцієнти стали менші за абсолютним значенням
  - Хоча картина все одно схожа
  - Тобто ми все одно можемо побачити факт *можливої* дискримінації
- Проте **некоректно** казати, що, скажімо, різниця в середніх зарплатах дорівнює -0.245 **винятково** за рахунок стати
  - Ми знаємо, що справжній вплив стати випливає з моделі на попередньому слайді
  - Умова про нульове умовне сподівання похибки вже не виконується

Ілюстративні приклади (3)

- Розгляньмо тепер вплив освіти на зарплату: $Y = m(X) + e$, $\mathbb{E}[e \mid X] = 0$
- Зрозуміло, що «насправді» навряд чи в цій моделі $\mathbb{E}[e \mid X] = 0$
  - Наприклад, можна очікувати, що e містить у собі стать, расу тощо
  - І тому **якщо** вони **корелюють** із X, то $\mathbb{E}[e \mid X] \neq 0$
  - Але для ілюстративних цілей **уявімо**, що окрім X на зарплату нічого більше в принципі не впливає
- Можливою лінійною апроксимацією $m(\mathbf{X})$ може бути лінійна функція з одним регресором X:

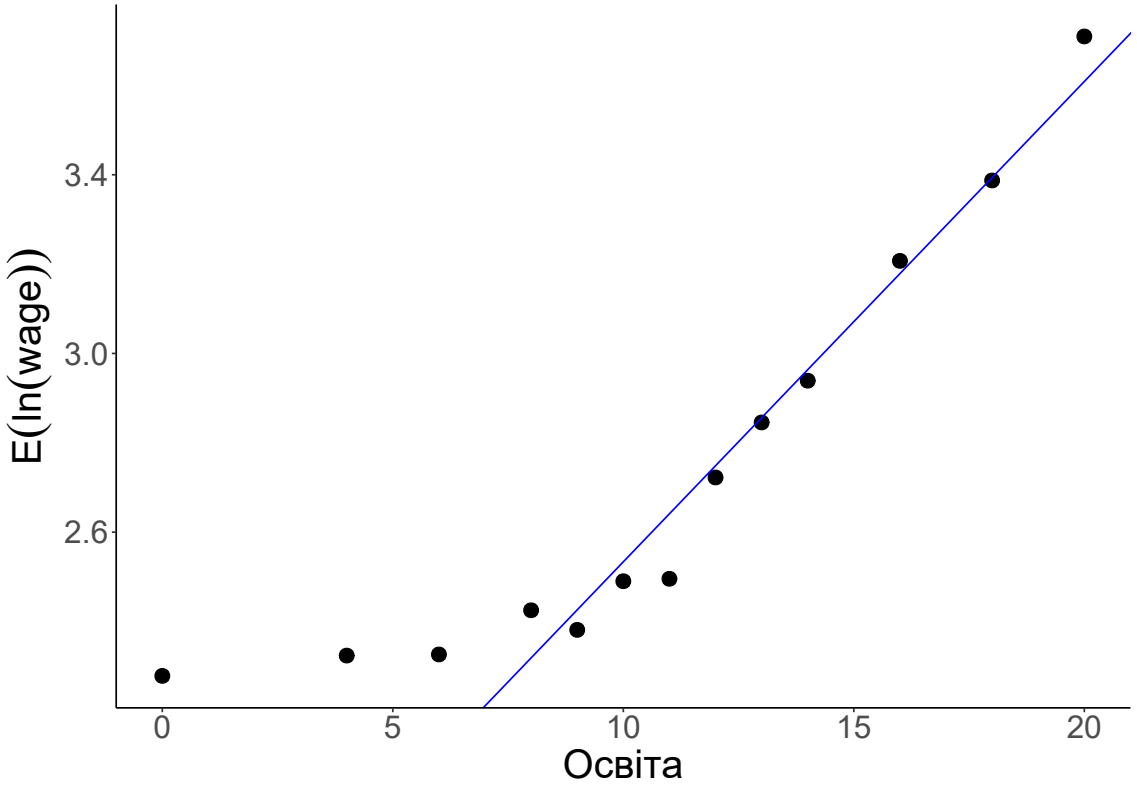
$$Y = \beta_0 + \beta_1 X + e$$

- За такого припущення можемо оцінити коефіцієнти цієї моделі як коефіцієнти лінійної проєкції:

$$\mathcal{P}(Y \mid X) = 1.458 + 0.108X$$

Ілюстративні приклади (4)

- Можемо намалювати, що вийшло:



- Як можна бачити, навряд чи можна вважати цю лінійну проєкцію доброю апроксимацією для CEF для Y
- Точніше, вона непогана тільки для освіти більше 9 років

Ілюстративні приклади (5)

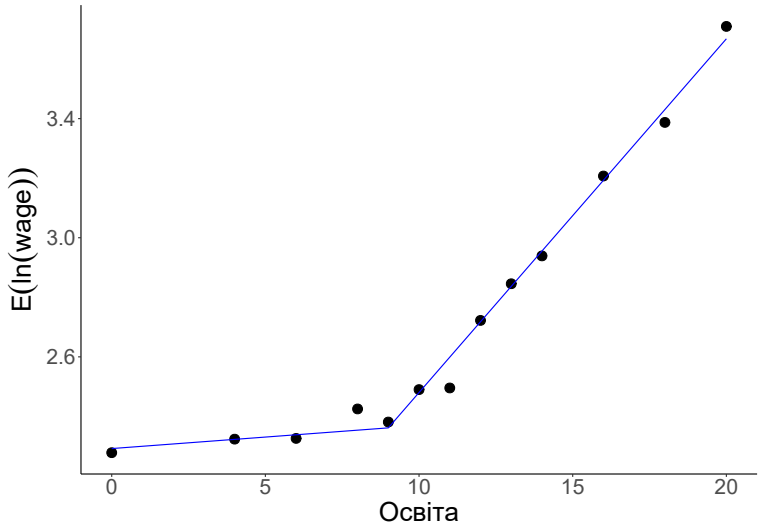
- Можна поліпшити відповідний результат, розглянувши таку модель:

$$Y = \beta_0 + \beta_1 X_1 + \beta_2 X_2 + e\ , \qquad X_1 = X\ ,\ X_2 = (X_1 - 9)\mathbb{1}\{X_1 > 9\}$$

- Тоді матимемо такі коефіцієнти лінійної регресії:

$$\mathcal{P}(Y \mid X_1, X_2) = 2.292 + 0.008 X_1 + 0.111 X_2$$

- Можемо намалювати, що вийшло:



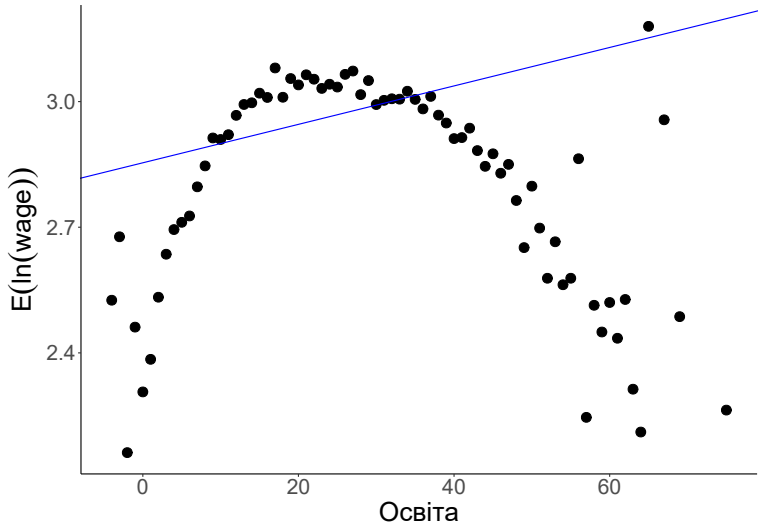
- Результат поліпшився

Ілюстративні приклади (6)

- Розгляньмо залежність зарплат від досвіду роботи X, який визначмо як вік за вирахуванням освіти та 6 (роки до походу в дитячий садок)
  - Знову вважатимемо, що ця змінна єдина, яка може впливати на зарплату
- Матимемо такий результат:

$$\mathcal{P}(Y \mid X) = 2.854 + 0.005X$$

- Можемо намалювати, що вийшло:



- Знову ж таки результат далекий від ідеального

Ілюстративні приклади (7)

- Але можна помітити, що залежність схожа на квадратичну
- Тому можна розглянути таку лінійну регресію:

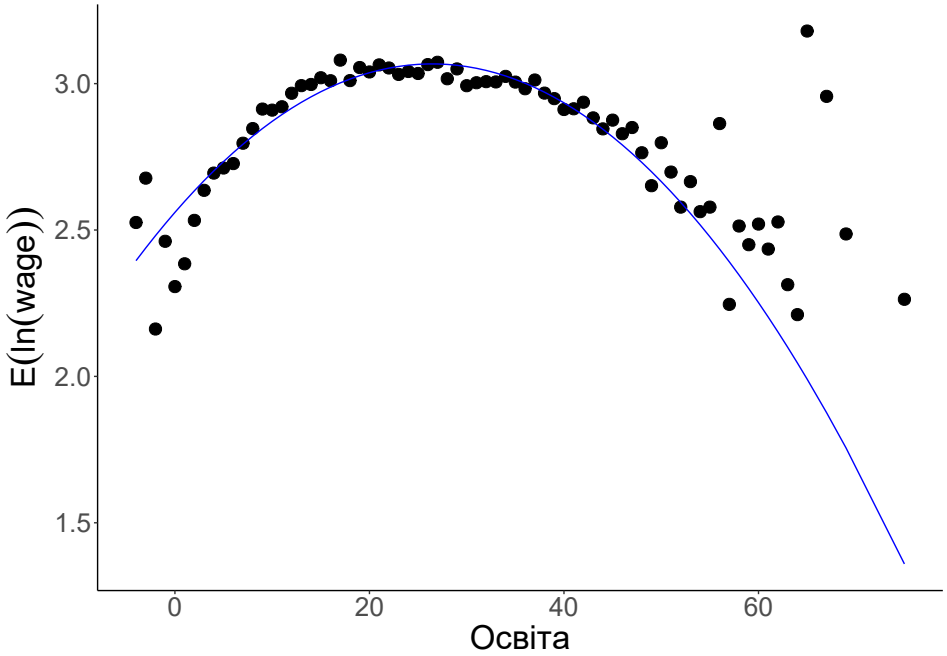
$$Y = \beta_0 + \beta_1 X + \beta_2 X^2 + e$$

- Матимемо такий результат:

$$\mathcal{P}(Y \mid X, X^2) = 2.5602 + 0.0383X - 7 \times 10^{-4} X^2$$

Ілюстративні приклади (8)

- Можемо намалювати, що вийшло:



- Результат значно ліпший
  - Можна, звісно, додавати ступені вищих порядків
  - Тоді апроксимація буде ще ліпшою
  - Але всьому є межа — ми не хочемо, щоб мало місце **перенавчання** (overfitting)

Різниця між структурними моделями та проєкціями (1)

- Вищенаведена дискусія проливає світло на питання якости лінійної апроксимації справжньої CEF
- Справжні проблеми з моделями на попередніх слайдах виникають, коли ми повертаємось у реальний світ
  - На практиці існує маса змінних, які корелюють зі статтю, расою, освітою і впливають на зарплату
  - Наприклад, вроджені здібності
- Тоді, наприклад, «справжня» модель повинна бути $Y = m(X_1, X_2, X_3, X_4) + e$, $\mathbb{E}[e \mid X_1, X_2, X_3, X_4] = 0$ (тобто m є CEF)
  - Тут X_1 — стать, X_2 — раса, X_3 — освіта, X_4 — вроджені здібності
- Але ми не можемо обчислити вроджені здібності, тому таких даних у нас просто немає
- Відповідно, ми **вимушені** оцінювати модель $Y = \beta_0 + \beta_1 X_1 + \beta_2 X_2 + \beta_3 X_3 + \tilde{e}$
  - Тобто це не лінійна апроксимація справжньої CEF — це просто неправильна модель
  - Можна очікувати, що $\text{Cov}(X_3, X_4) > 0$ (хто має більше здібностей, у того вища освіта)
  - Отже $\text{Cov}(X_3, \tilde{e}) \neq 0$, бо $\text{Cov}(X_3, X_4) \neq 0$, а $\tilde{e}$ «містить у собі» X_4
  - Тому вплив освіти на зарплату вже не буде ізольованим
  - І можна поставити ці результати під сумнів, стверджуючи, що це вроджені здібності мають вплив, а освіта мало на що впливає
  - Максимум, про що можна говорити — про певний **статистичний зв'язок** між змінними

Різниця між структурними моделями та проєкціями (2)

- Оцінити цю некоректну модель ми можемо як лінійну проєкцію $Y = \gamma_0 + \gamma_1 X_1 + \gamma_2 X_2 + \gamma_3 X_3 + u$
  - R порахує нам усе, що завгодно!
- І за властивостями лінійної проєкції буде автоматично виконуватися $\mathbb{E}[X_i u] = 0$
  - Але $u \neq \tilde{e}$, а $\beta_i \neq \gamma_i$!
  - Тому ми порахували щось зовсім інше
- Модель $Y = \beta_0 + \beta_1 X_1 + \beta_2 X_2 + \beta_3 X_3 + \tilde{e}$ називають **структурною** (structural)
  - У ній ми **нічого не знаємо** про $\tilde{e}$ і її корельованість із регресорами
  - Ми просто **хочемо** оцінити $\mathbb{E}[Y \mid X_1, X_2, X_3]$, і інші змінні нас особливо не цікавлять
  - Але вони нас можуть не цікавити, але **псувати нам оцінку**
- Модель $Y = \gamma_0 + \gamma_1 X_1 + \gamma_2 X_2 + \gamma_3 X_3 + u$, де $\mathbb{E}[u\mathbf{X}] = 0$, є **лінійною проєкцією**
  - Її коефіцієнти **не будуть** дорівнювати коефіцієнтам структурної моделі, якщо $\mathbb{E}[\tilde{e}\mathbf{X}] \neq 0$
- Тобто записати можна все, що завгодно, і оцінити як лінійну проєкцію також, але чи буде це відповідати нашим потребам — велике питання
  - У цьому полягає складність застосування регресійного аналізу та методів причиново-наслідкового виведення
  - Ми до цього повертатимемося постійно, і особливу увагу звернемо наприкінці нашого курсу
  - А поки що розглянемо статистичні властивості оцінки коефіцієнтів лінійної проєкції

План лекції

Plug-in оцінка (1)

- Як ми зазначали в попередніх лекціях, простим способом дістати оцінку деякого параметра є замінити всі теоретичні моменти на емпіричні
- Нехай маємо вибірку $(Y_i, X_{1,i}, \dots, X_{k,i}), i = 1, \dots, n$
  - Вважаємо, що всі n векторів незалежні **між собою**
  - ...та мають однаковий розподіл $\mathbb{P}_{Y,X_1,\dots,X_k}$, який нам невідомий
- Нехай відповідні випадкові величини пов'язані лінійною залежністю $Y_i = \beta_0 + \sum_{i=1}^{k} \beta_i X_i + u_i$
  - Тут $\mathbb{E}[X_i u_i] = 0$, $\mathbb{E}[u_i] = 0$ суто за побудовою, бо це лінійна проєкція
  - Ми нічого не кажемо про якийсь причиново-наслідковий чи інший зміст цієї моделі
- Позначмо $\mathbf{X}_i = (1, X_{1,i}, \dots, X_{k,i})^\top$
- Тоді відповідні коефіцієнти проєкції за (2.3) дорівнюють

$$\beta = \mathbf{Q}_{\mathbf{XX}}^{-1} \mathbf{Q}_{\mathbf{X}Y} = \left(\mathbb{E}\left[\mathbf{X}_i \mathbf{X}_i^\top\right]\right)^{-1} \mathbb{E}\left[\mathbf{X}_i Y_i\right]$$

- Відтак plug-in оцінкою цих коефіцієнтів буде

$$\hat{\beta} = \left(\frac{1}{n}\sum_{i=1}^{n} \mathbf{X}_i \mathbf{X}_i^\top\right)^{-1} \left(\frac{1}{n}\sum_{i=1}^{n} \mathbf{X}_i Y_i\right) \tag{3.1}$$

Plug-in оцінка (2)

- Альтернативний запис можна дістати, записавши всі рівняння як матричні:

$$\mathbf{Y} = \underset{n\times 1}{\begin{pmatrix} Y_1 \\ Y_2 \\ \vdots \\ Y_n \end{pmatrix}}, \quad \mathbf{X} = \underset{n\times(k+1)}{\begin{pmatrix} 1 & X_{1,1} & \dots & X_{k,1} \\ 1 & X_{1,2} & \dots & X_{k,2} \\ \vdots & \vdots & \ddots & \vdots \\ 1 & X_{1,n} & \dots & X_{k,n} \end{pmatrix}} \equiv \begin{pmatrix} \mathbf{X}_1^\top \\ \mathbf{X}_2^\top \\ \vdots \\ \mathbf{X}_n^\top \end{pmatrix}$$

- Тоді нашу лінійну модель можна записати в матричній формі як

$$\mathbf{Y} = \mathbf{X}\beta + \mathbf{u}\,, \tag{3.2}$$

де $\mathbf{u} = (u_1, u_2, \dots, u_n)^\top$

- І тоді оцінка β дорівнюватиме

$$\hat{\beta} = \left(\frac{1}{n}\sum_{i=1}^{n}\mathbf{X}_i\mathbf{X}_i^\top\right)^{-1}\left(\frac{1}{n}\sum_{i=1}^{n}\mathbf{X}_i Y_i\right) = \left(\mathbf{X}^\top\mathbf{X}\right)^{-1}\mathbf{X}^\top\mathbf{Y}$$

Метод найменших квадратів (1)

- Виявляється, цей же результат можна дістати зовсім з інших міркувань
- Можемо розглянути задачу мінімізації **суми квадратів похибок** (sum of squared errors, SSE) $u_i = Y_i - \mathbf{X}_i^\top \beta$ для всіх $i = 1, \dots, n$
- У матричній формі SSE $= \mathbf{u}^\top \mathbf{u}$, відтак маємо таку задачу мінімізації:

$$\begin{aligned}\hat{\beta} = \arg\min_{\mathbf{b}} \text{SSE}(\mathbf{b}) &= \arg\min_{\mathbf{b}} \mathbf{u}^\top \mathbf{u} \\ &= \arg\min_{\mathbf{b}} (\mathbf{Y} - \mathbf{X}\mathbf{b})^\top (\mathbf{Y} - \mathbf{X}\mathbf{b}) \\ &= \arg\min_{\mathbf{b}} (\mathbf{Y}^\top \mathbf{Y} - \mathbf{Y}^\top \mathbf{X}\mathbf{b} - \mathbf{b}^\top \mathbf{X}^\top \mathbf{Y} + \mathbf{b}^\top \mathbf{X}^\top \mathbf{X}\mathbf{b})\end{aligned}$$

  - Зверніть увагу, що $\mathbf{Y}^\top \mathbf{X}\mathbf{b} = \mathbf{b}^\top \mathbf{X}^\top \mathbf{Y}$, оскільки обидва є скаляри

Метод найменших квадратів (2)

- Умови першого порядку дорівнюють

$$\frac{\partial \text{SSR}}{\partial \mathbf{b}} = -2\left(\mathbf{X}^\top \mathbf{Y}\right) + 2\left(\mathbf{X}^\top \mathbf{X}\right)\mathbf{b} = 0$$

- Звідси

$$\left(\mathbf{X}^\top \mathbf{X}\right)\mathbf{b} = \mathbf{X}^\top \mathbf{Y}$$

- Якщо $\mathbf{X}^\top \mathbf{X}$ має обернену[5], то остаточно маємо

$$\hat{\beta} = \left(\mathbf{X}^\top \mathbf{X}\right)^{-1}\left(\mathbf{X}^\top \mathbf{Y}\right)$$

  - Це те саме, що й (3.1)
  - Умову другого порядку можете перевірити самостійно
- Із цих міркувань оцінку $\hat{\beta}$ називають **оцінкою найменших квадратів** (ordinary least squares estimator, OLS)

[5] Конкретніше, якщо ця матриця додатно визначена

Алгебричні властивості залишків

- Після того, як ми оцінили коефіцієнти регресії $\hat{\beta}$, можна обчислити **прогнозні значення** (fitted values) залежної змінної $\hat{Y}_i = \mathbf{X}_i^\top \hat{\beta}$
- **Залишки** (residuals) є різниці між справжніми та прогнозними значеннями:

$$\hat{u}_i = Y_i - \hat{Y}_i = Y_i - \mathbf{X}_i^\top \hat{\beta}$$

- З умов першого порядку маємо

$$\left(\mathbf{X}^\top \mathbf{X}\right) \hat{\beta} = \mathbf{X}^\top \mathbf{Y}$$

$$\mathbf{X}^\top \left(\mathbf{Y} - \mathbf{X}\hat{\beta}\right) = \mathbf{0}$$

$$\mathbf{X}^\top \left(\hat{u}_1, \dots, \hat{u}_n\right)^\top = \mathbf{0}$$

- Оскільки перший рядок $\mathbf{X}^\top$ містить самі одиниці,

$$\sum_{i=1}^{n} \hat{u}_i = 0 \tag{3.3}$$

  - Тобто *вибіркове середнє* залишків дорівнює 0
- Також, оскільки j-ий рядок $\mathbf{X}^\top$ дорівнює $(X_{j,1}, \dots, X_{j,n})$,

$$\sum_{i=1}^{n} X_{j,i} \hat{u}_i = 0 \tag{3.4}$$

  - Тобто *вибіркова коваріація* між залишками і будь-яким регресором дорівнює 0

Коефіцієнт детермінації R^2 (1)

- Ми можемо оцінити, наскільки доброю є наша оцінка регресійної моделі, використовуючи такі величини
- **Сума квадратів похибок** $\text{SSE} = \sum_{i=1}^{n} \hat{u}_i^2 = \sum_{i=1}^{n} \left(Y_i - \hat{Y}_i\right)^2$
  - Ми її мінімізували, щоб дістати $\hat{\beta}$
- **Пояснена сума квадратів** (explained sum of squares) $\text{ESS} = \sum_{i=1}^{n} \left(\hat{Y}_i - \overline{Y}\right)^2$
  - Фактично, це дисперсія $\hat{Y}_i$, адже середнє $\hat{Y}_i$ дорівнює $\overline{Y}$
  - Оскільки $\hat{Y}_i$ є функцією від $\mathbf{X}_i$, це дисперсія, пов'язана з $\mathbf{X}_i$
- **Повна сума квадратів** (total sum of squares) $\text{TSS} = \sum_{i=1}^{n} \left(Y_i - \overline{Y}\right)^2$
  - Фактично, це дисперсія Y_i
- Можна показати[6], що $\text{TSS} = \text{ESS} + \text{SSE}$
  - Зокрема, якщо $\text{TSS} = \text{ESS}$, регресійна гіперплощина проходить через усі точки набору даних

[6] Це нескладно, але довго й нудно

Коефіцієнт детермінації R^2 (2)

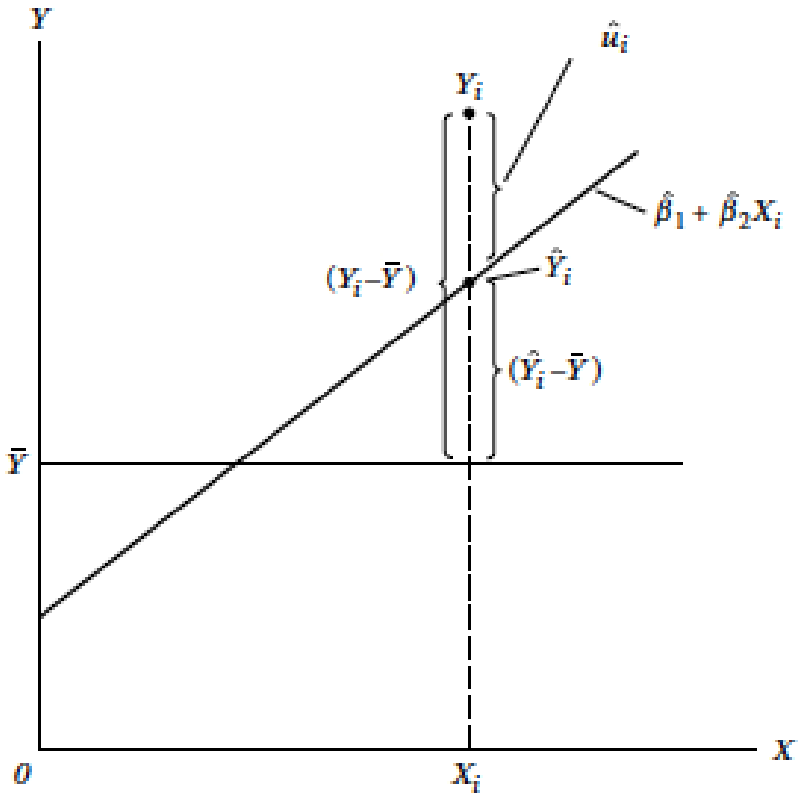
- Можна розглянути поняття **коефіцієнта детермінації**:

$$R^2 = \frac{\text{ESS}}{\text{TSS}} = 1 - \frac{\text{SSE}}{\text{TSS}} \tag{3.5}$$

- R^2 показує, яку частку дисперсії Y_i пояснює дисперсія в $\mathbf{X}_i$
  - Іншими словами, наскільки якісно прогнозні значення $\hat{Y}_i$ збігаються з даними
  - $0 \leq R^2 \leq 1$ за побудовою
  - $R^2 = 1$ відповідає ситуації, коли залишки повністю нульові
- За визначенням, R^2 вищий тоді, коли
  - Залишки мають меншу дисперсію (розкид залишків навколо регресійної площини малий)
  - Дисперсія Y_i вища

Коефіцієнт детермінації R^2 (3)

- Графічна ілюстрація:



- Як можна бачити, що менше відношення SSE до TSS, то ліпше наближення
- Відтак часто R^2 використовують як міру якости лінійної модели
- Проте варто пам'ятати, що з точки зору причиново-наслідкового аналізу R^2 **не має жодного значення**!

Коефіцієнт детермінації R^2 (4)

- Потрібно помітити, що R^2 зі збільшенням кількости незалежних змінних може тільки рости
- Оцінка OLS за побудовою мінімізує SSE, а відтак додавання нових змінних змушує нас шукати мінімум у просторі більшої розмірности
  - Відтак значення SSE ніяк не може зрости
- Трішки формальніше, модель із k регресорами можна розглядати як модель із $k+1$ регресорами, де $\beta_{k+1} = 0$
- Тоді оптимізаційна задача $\arg\min_{b_0,b_1,\dots,b_k} \text{SSE}(b_0, b_1, \dots, b_k)$ еквівалентна задачі $\arg\min_{b_0,b_1,\dots,b_k,b_{k+1}} \text{SSE}(b_0, b_1, \dots, b_k, 0)$
- Ця задача є задачею **умовної** оптимізації, а відтак її відповідь не може бути ліпшою, ніж у задачі безумовної:

$$\min_{b_0,b_1,\dots,b_k,b_{k+1}} \text{SSE}(b_0, b_1, \dots, b_k, 0) \geq \min_{b_0,b_1,\dots,b_k,b_{k+1}} \text{SSE}(b_0, b_1, \dots, b_k, b_{k+1})$$

Коефіцієнт детермінації R^2 (5)

- Це не дуже добре, адже це означає, що R^2 можна збільшувати, просто додаючи будь-який мотлох у модель
- Натомість на практиці ліпше використовувати **скоригований** (adjusted) R^2:

$$R^2_{\text{adj}} = 1 - \frac{\text{SSE}/(n-k-1)}{\text{TSS}/(n-1)} \tag{3.6}$$

  - Ми фактично всюди додали корекції зміщення з урахуванням ступенів свободи
- Потрібно помітити, що $R^2_{\text{adj}} < R^2$, оскільки $\frac{n-1}{n-k-1} > 1$
  - R^2_{adj} навіть може набувати від'ємних значень
- Знову варто наголосити, що для причиново-наслідкового аналізу R^2_{adj} **зовсім** не є корисним:
  - Збільшення R^2_{adj} **не означає**, що новододана змінна статистично значуща
  - Високі значення R^2_{adj} **не означають**, що існує будь-який причиново-наслідковий зв'язок між змінними
  - Високі значення R^2_{adj} **не означають**, що лінійна модель є коректна

Стандартна похибка регресії

- Також можна розглянути **стандартну похибку регресії** (standard error of the regression, SER)
- Це просто оцінка середньоквадратичного відхилення похибок
  $\sigma_u = \sqrt{\text{Var}\left(u_i\right)} = \sqrt{\mathbb{E}\left[u_i^2\right] - \left(\mathbb{E}\left[u_i\right]\right)^2} = \sqrt{\mathbb{E}\left[u_i^2\right]}$
- Ми не спостерігаємо справжніх u_i, а тільки маємо доступ до залишків $\hat{u}_i$
- Відтак стандартною похибкою регресії є величина

$$\text{SER} = \hat{\sigma}_u\ , \quad \hat{\sigma}_u^2 = \frac{1}{n-k-1}\sum_{i=1}^{n}\hat{u}_i^2 = \frac{\text{SSR}}{n-k-1} \tag{3.7}$$

  - Ми ділимо на $n-k-1$ замість n зі схожих міркувань, чому ми ділимо на $n-1$ замість n, обчислюючи вибіркову дисперсію (корекція на ступені вільності)
  - Така оцінка буде спроможною
  - Також вона буде незміщеною, але тільки якщо дисперсія u_i не залежить від $\mathbf{X}_i$
  - Ми на цьому зупинятися сильно не будемо
- У будь-якому випадку SER показує розкид даних навколо регресійної площини

Гетероскедастичність та гомоскедастичність

- Якщо $\text{Var}\left(u_i \mid \mathbf{X}_i\right) = \sigma_u^2$, тобто якщо дисперсія u_i **не залежить** від $\mathbf{X}_i$, то кажуть, що похибки **гомоскедастичні** (homoskedastic)
- В іншому випадку кажуть, що вони **гетероскедастичні** (heteroskedastic)
- З історичних причин завжди згадують про гомоскедастичні похибки
  - Якщо це виконується, то можна довести деякі теоретичні властивості OLS оцінок
  - Коли не було доступу до потужних комп'ютерів, такі припущення спрощували обчислення та аналіз
- На практиці **майже завжди** зустрічаються тільки гетероскедастичні похибки
  - На жаль, майже в усіх стандартних програмних пакетах за замовчуванням рахують гомоскедастичні похибки
  - Ми до цього ще повернемося
- Простий приклад: $Y_i = \beta_0 + \beta_1 X_i + u_i$, де Y_i — витрати на їжу, X_i — дохід людини
  - Можна очікувати, що $\beta_1 > 0$
  - Також можна очікувати, що дисперсія Y_i буде вищою для осіб з вищим X_i
  - Незаможні люди майже всі кошти витрачають на їжу
  - Що більше стає дохід людини, то більше свободи вона має щодо розподілу свого бюджету
  - Тому дисперсія u_i буде змінюватися (зростати) з X_i

OLS оцінка в R (1)

- Для того, щоб зробити оцінювання за методом OLS та провести аналіз результатів, в R існує функція `lm` (від linear model)
- Розгляньмо для прикладу дані про результати тестування з читання та математики учнів 5-х класів шкіл Каліфорнії (1999 р.)[7]

```
caschool <- read_csv("data/caschool.csv")
```

```
##
## -- Column specification --------------------------------------------------------
## cols(
##   observation_number = col_double(),
##   dist_cod = col_double(),
##   county = col_character(),
##   district = col_character(),
##   gr_span = col_character(),
##   enrl_tot = col_double(),
##   teachers = col_double(),
##   calw_pct = col_double(),
##   meal_pct = col_double(),
##   computer = col_double(),
##   testscr = col_double(),
##   comp_stu = col_double(),
##   expn_stu = col_double(),
##   str = col_double(),
##   avginc = col_double(),
##   el_pct = col_double(),
##   read_scr = col_double(),
##   math_scr = col_double()
## )
```

[7] Приклад узято з книжки Stock & Watson

OLS оцінка в R (2)

- Можемо обчислити коефіцієнти моделі $Y_i = \beta_0 + \beta_1 X_i + u_i$
  - Кожний i відповідає окремому шкільному округу
  - X_i відповідає середньому числу учнів на одного вчителя `str`
  - Y_i відповідає середньому балу за тест в окрузі `testscr`
- У функції `lm` потрібно зліва від ~ вказати залежну змінну, а справа — всі незалежні (константа включається за замовчуванням)

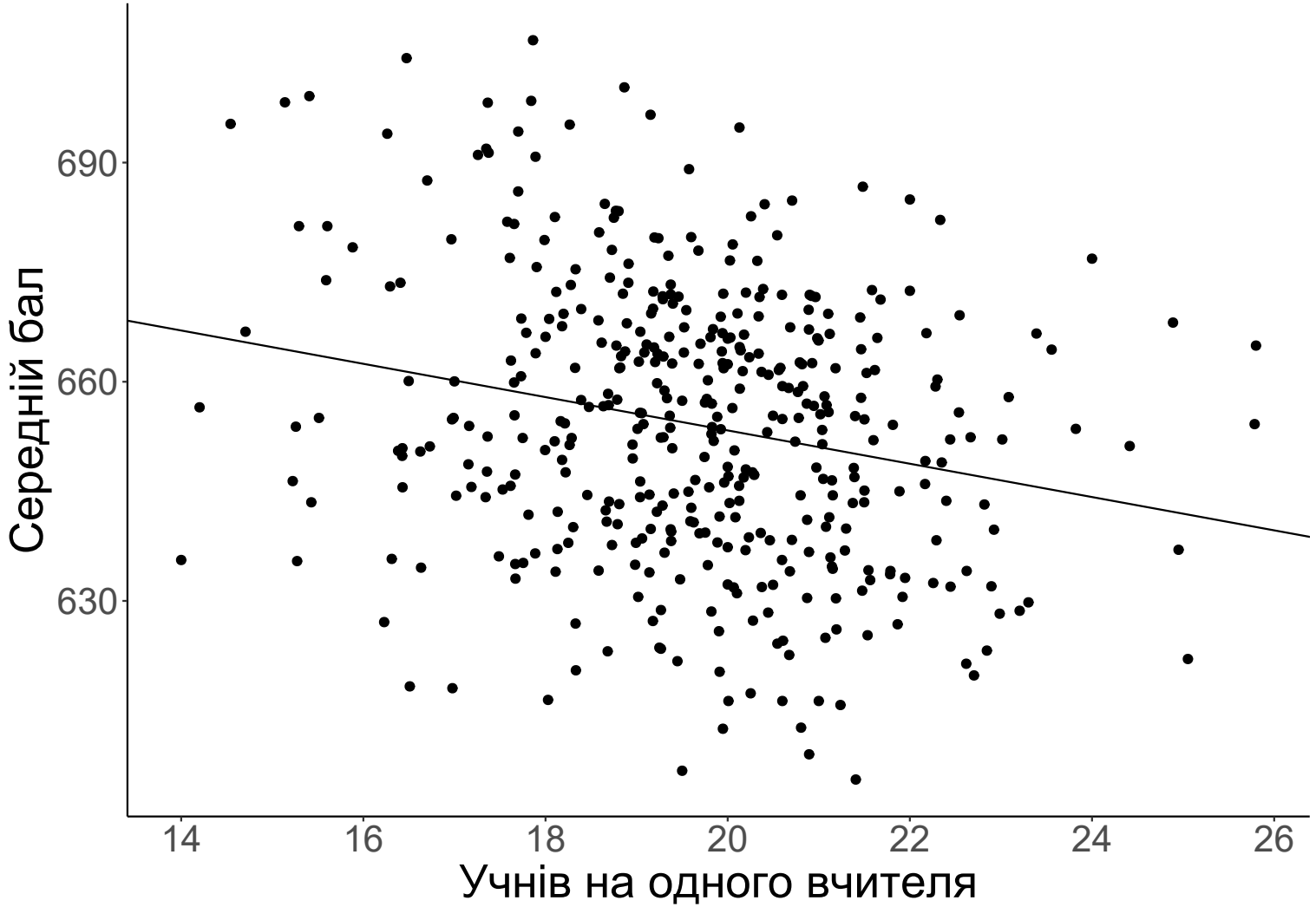
```
model <- lm(testscr ~ str, data = caschool)
model
```

```
##
## Call:
## lm(formula = testscr ~ str, data = caschool)
##
## Coefficients:
## (Intercept)          str
##      698.93        -2.28
```

- Бачимо, що збільшення числа учнів, що припадає на одного вчителя, на 1, *пов'язано* зі зменшенням середнього балу за тести на -2.28
  - Ми поки що не знаємо, чи є це статистично значущим (можливо, це 0)!
    - Це предмет наступної лекції
  - Ми не можемо інтерпретувати це як причиново-наслідковий зв'язок
    - Тому що в нас існують об'єктивні сумнів, що u_i некорельована з X_i
    - Існують інші фактори, які впливають на бали за тести і корелюють із розміром класу

OLS оцінка в R (3)

- Графічна ілюстрація



OLS оцінка в R (4)

- Більше додаткової інформації про модель можна дістати, застосувавши функцію `summary`

```
summary(model)
```

```
##
## Call:
## lm(formula = testscr ~ str, data = caschool)
##
## 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) 698.9330     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
```

- Тут дуже багато корисної інформації, яка нам знадобиться наступної лекції
- Поки що можемо помітити, що $R^2 = 0.051$, $R^2_{\text{adj}} = 0.049$
- Їх можна порахувати самостійно, адже `model` містить як прогнозні значення, так і залишки:

```
n <- nrow(caschool)
1 - mean(model$residuals^2) / mean((caschool$testscr - mean(caschool$testscr))^2)
```

```
## [1] 0.0512401
```

```
1 - (sum(model$residuals^2) / (n - 2)) / (sum((caschool$testscr - mean(caschool$testscr))^2) / (n - 1))
```

```
## [1] 0.04897034
```

План лекції

Припущення

- Розгляньмо основні властивості, які має OLS оцінка коефіцієнтів лінійної регресії
- Для того, щоб їх вивести, потрібно зробити декілька **припущень**
- **Припущення 1**: умовне сподівання похибок дорівнює 0: $\mathbb{E}\left[u_i \mid \mathbf{X}_i\right] = 0$
  - Без цього припущення ми не зможемо вважати, що наша лінійна модель є (хоча б апроксимацією) CEF
  - Відтак жодної, навіть натягнутої, причиново-наслідкової інтерпретації β не матимуть
- **Припущення 2**: $(X_{1,i}, \dots, X_{k,i}, Y_i)$, $i = 1, \dots, n$ незалежні та мають однаковий розподіл
- **Припущення 3**: $0 < \mathbb{E}\left[X_{j,i}^4\right] < \infty\ ,\ j = 1, \dots, k\ , \quad 0 < \mathbb{E}\left[Y_i^4\right] < \infty$
  - Фактично, мається на увазі, що в даних немає великих викидів
  - Формально математична роль цього припущення полягає в тому, що можна буде застосувати ЗВЧ для вибіркових дисперсій
- **Припущення 4**: матриця $\mathbf{X}^\top \mathbf{X}$ має обернену
  - Іншими словами, $\mathbf{X}_{n \times (k+1)}$ має ранг $k + 1$
  - Якщо це не виконуватиметься, то неможливо буде порахувати (єдину) OLS оцінку
  - На практиці це припущення порушується, якщо деякі змінні є лінійними комбінаціями інших
  - Наприклад, наявність одночасно і рівня освіти, і досвіду роботи

Незміщеність

- Можемо показати, що OLS оцінка є незміщеною
- Можна розписати $\mathbf{Y}$ і дістати

$$\hat{\beta} = \left(\mathbf{X}^\top \mathbf{X}\right)^{-1} \mathbf{X}^\top \left(\mathbf{X}\beta + \mathbf{u}\right) = \left(\mathbf{X}^\top \mathbf{X}\right)^{-1} \mathbf{X}^\top \mathbf{X}\beta + \left(\mathbf{X}^\top \mathbf{X}\right)^{-1} \mathbf{X}^\top \mathbf{u}$$

- Добуток матриці на обернену дає одиничну матрицю, тому

$$\hat{\beta} = \beta + \left(\mathbf{X}^\top \mathbf{X}\right)^{-1} \mathbf{X}^\top \mathbf{u}$$

- Звідси

$$\mathbb{E}\left[\hat{\beta} \mid \mathbf{X}\right] = \beta + \left(\mathbf{X}^\top \mathbf{X}\right)^{-1} \mathbf{X}^\top \mathbb{E}\left[\mathbf{u} \mid \mathbf{X}\right]$$

- Останній доданок дорівнює 0 за Припущеннями 1 і 2:

$$\mathbb{E}\left[\mathbf{u} \mid \mathbf{X}\right] = \begin{pmatrix} \mathbb{E}\left[u_1 \mid \mathbf{X}\right] \\ \vdots \\ \mathbb{E}\left[u_n \mid \mathbf{X}\right] \end{pmatrix} = \begin{pmatrix} \mathbb{E}\left[u_1 \mid \mathbf{X}_1\right] \\ \vdots \\ \mathbb{E}\left[u_n \mid \mathbf{X}_n\right] \end{pmatrix} = \mathbf{0}$$

- За законом ітерованих сподівань також випливає, що $\mathbb{E}\left[\hat{\beta}\right] = \mathbb{E}\left[\mathbb{E}\left[\hat{\beta} \mid \mathbf{X}\right]\right] = \beta$

Спроможність

- Розпишімо $\mathbf{Y}$, як і на попередньому слайді, але використаймо версію формули з сумами:

$$\hat{\beta} = \beta + \left(\frac{1}{n} \sum_{i=1}^{n} \mathbf{X}_i \mathbf{X}_i^\top \right)^{-1} \left(\frac{1}{n} \sum_{i=1}^{n} \mathbf{X}_i u_i \right)$$

- Згідно з ЗВЧ,

$$\frac{1}{n} \sum_{i=1}^{n} \mathbf{X}_i \mathbf{X}_i^\top \xrightarrow{p} \mathbf{Q}_{\mathbf{XX}}$$

$$\frac{1}{n} \sum_{i=1}^{n} \mathbf{X}_i u_i \xrightarrow{p} \mathbb{E}\left[\mathbf{X}_i u_i\right]$$

- Згідно з ТНВ (усі функції неперервні, а за Припущенням 4 матриця має обернену),

$$\left(\frac{1}{n} \sum_{i=1}^{n} \mathbf{X}_i \mathbf{X}_i^\top \right)^{-1} \left(\frac{1}{n} \sum_{i=1}^{n} \mathbf{X}_i u_i \right) \xrightarrow{p} \mathbf{Q}_{\mathbf{XX}}^{-1} \mathbb{E}\left[\mathbf{X}_i u_i\right]$$

- За Припущенням 1 та законом ітерованих сподівань маємо $\mathbb{E}\left[\mathbf{X}_i u_i\right] = \mathbf{0}$
- Відтак $\hat{\beta} \xrightarrow{p} \beta$

Асимптотичний розподіл

- Якщо перенести β вліво та помножити на $\sqrt{n}$, дістанемо:

$$\sqrt{n}\left(\hat{\beta}-\beta\right)=\left(\frac{1}{n}\sum_{i=1}^{n}\mathbf{X}_i\mathbf{X}_i^\top\right)^{-1}\left(\frac{1}{\sqrt{n}}\sum_{i=1}^{n}\mathbf{X}_i u_i\right)$$

- Оскільки $\mathbb{E}\left[\mathbf{X}_i u_i\right]=0$, а $\mathbf{X}_i u_i$, $i=1,\dots,n$, незалежні, можемо застосувати ЦГТ до другого множника:

$$\frac{1}{\sqrt{n}}\sum_{i=1}^{n}\mathbf{X}_i u_i=\sqrt{n}\left(\frac{1}{n}\sum_{i=1}^{n}\mathbf{X}_i u_i-\mathbf{0}\right)\xrightarrow{d}N\left(\mathbf{0},\operatorname{Var}\left(\mathbf{X}_i u_i\right)\right)=N\left(\mathbf{0},\mathbb{E}\left[\mathbf{X}_i\mathbf{X}_i^\top u_i^2\right]\right)$$

  - Припущення 3 гарантує скінченність відповідної дисперсії
- Тоді за теоремою Слуцького

$$\sqrt{n}\left(\hat{\beta}-\beta\right)\xrightarrow{d}N\left(0,\mathbf{Q}_{\mathbf{xx}}^{-1}\mathbb{E}\left[\mathbf{X}_i\mathbf{X}_i^\top u_i^2\right]\mathbf{Q}_{\mathbf{xx}}^{-1}\right)\equiv N\left(0,\mathbf{V}_\beta\right)\tag{4.1}$$

- Зверніть увагу, що асимптотична дисперсія залежить від **невідомих** u_i
  - Наступного разу дізнаємося, що з цим робити

Симуляція Монте-Карло (1)

- Проілюструймо статистичні властивості OLS за допомогою симуляції Монте-Карло
- Розгляньмо вигадану модель, подібну на розглянутий вище приклад із тестовими балами Y_i та числом учнів на одного вчителя X_i:

$$Y_i = 660 - 2X_i + u_i \ , \quad \mathbb{E}\left[u_i \mid X_i\right] = 0$$

  - Оскільки ми написали умову про нульове умовне сподівання, ми вважаємо, що це наша **справжня** CEF
  - Тут $\beta_0 = 660$ and $\beta_1 = -2$
- Також вважатимемо, що $X_i \sim N(20, 4)$
- А похибки $u_i \mid X_i \sim N\left(0, 0.25 \cdot (X_i - 15)^2\right)$
  - Ми моделюємо їх як гетероскедастичні похибки
  - Варто помітити, що за такого моделювання справді $\mathbb{E}\left[u_i \mid X_i\right] = 0$

Симуляція Монте-Карло (2)

- Функція для генерування однієї вибірки:

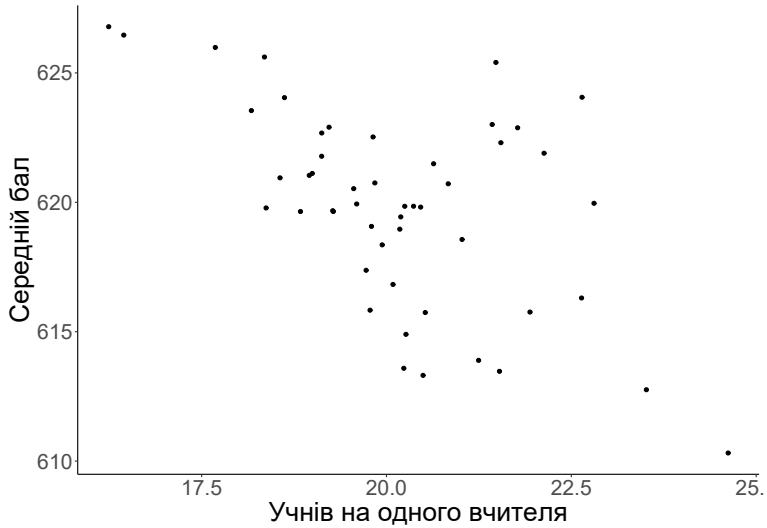
```
generate_X <- function(n){
  X <- rnorm(n, 20, sd = 2)
}

generate_data <- function(n, beta_0, beta_1, X = NULL){
  if (is.null(X)){
    X <- generate_X(n)
  }

  u <- rnorm(n, 0, sd = sqrt(0.25*(X - 15)^2))
  Y <- beta_0 + beta_1*X + u

  return(data.frame(Y, X, u))
}
```

- Результат для однієї вибірки:



Симуляція Монте-Карло (3)

- Ми зробимо три різні симуляції
- Для того, щоб показати, що OLS оцінка є *умовно* незміщена, $\mathbb{E}\left[\hat{\beta} \mid \mathbf{X}\right] = \beta$, ми згенеруємо $\mathbf{X}$ **один раз**
  - Потім ми $T = 5\,000$ разів утворюватимемо нові вибірки $(Y_i, X_i)^\top$, генеруючи щоразу нові u_i

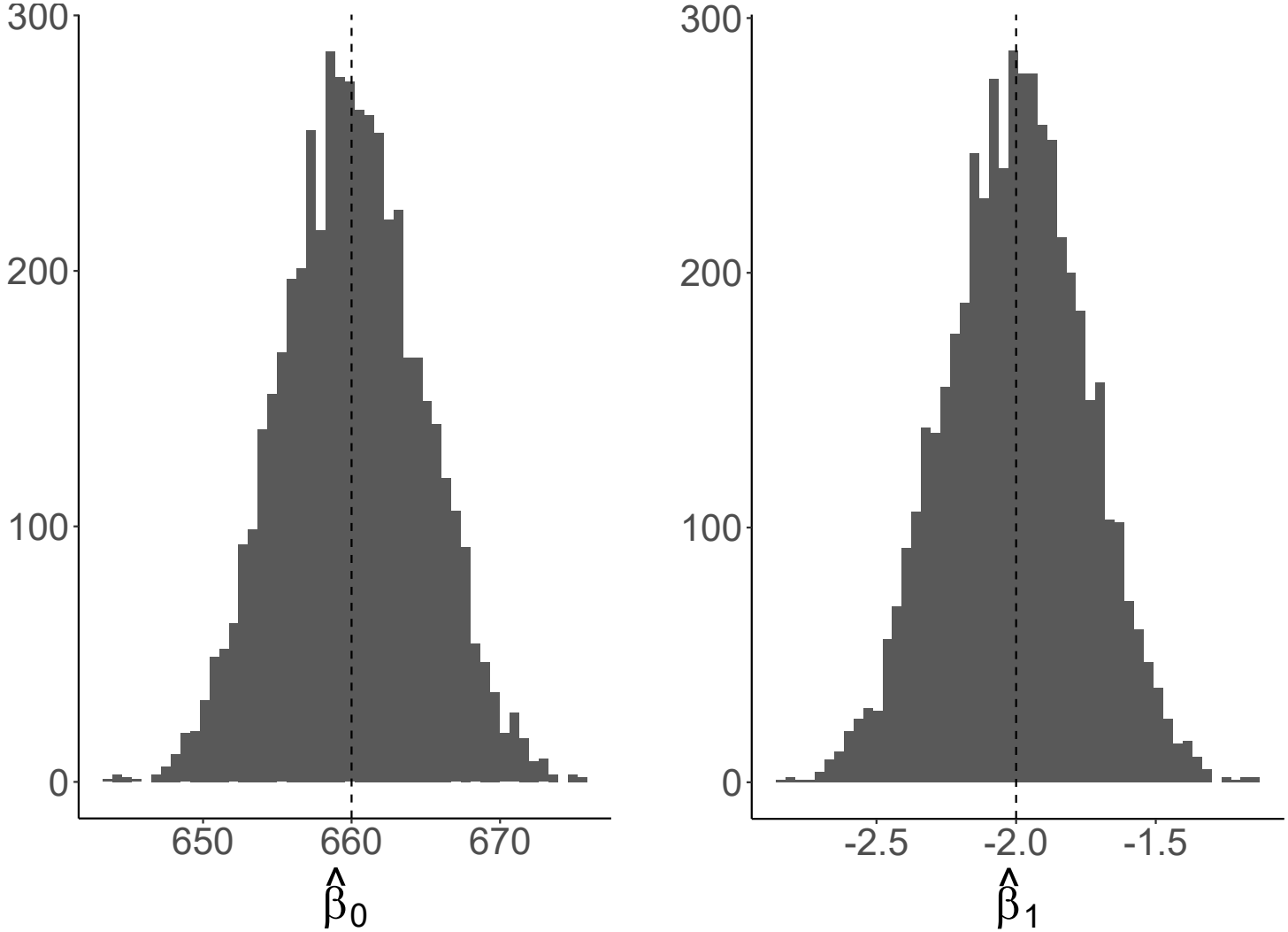
```
get_beta <- function(n, beta_0, beta_1, X){
  data <- generate_data(n, beta_0, beta_1, X)
  model <- lm(Y ~ X, data = data)
  return(model$coefficients)
}

T <- 5000
X <- generate_X(n)

df <- as_tibble(t(replicate(T, get_beta(n, beta_0, beta_1, X))))
```

Симуляція Монте-Карло (4)

- Відповідні гістограми:



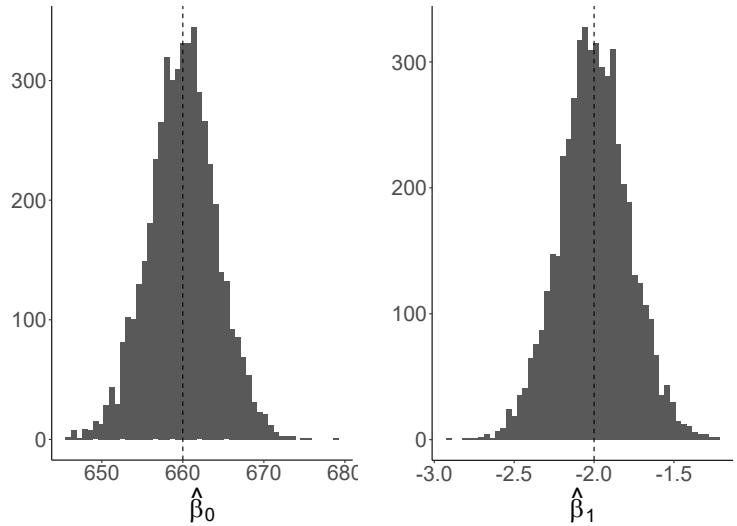
- Як можна бачити, оцінки справді незміщені

Симуляція Монте-Карло (5)

- Для того, щоб показати, що OLS оцінка є *безумовно* незміщена, $\mathbb{E}\left[\hat{\beta}\right] = \beta$, ми генеруватимемо **X наново** для кожної вибірки

```
df <- as_tibble(t(replicate(T, get_beta(n, beta_0, beta_1, X = NULL))))
```

- Відповідні гістограми:



- Як можна бачити, оцінки справді незміщені

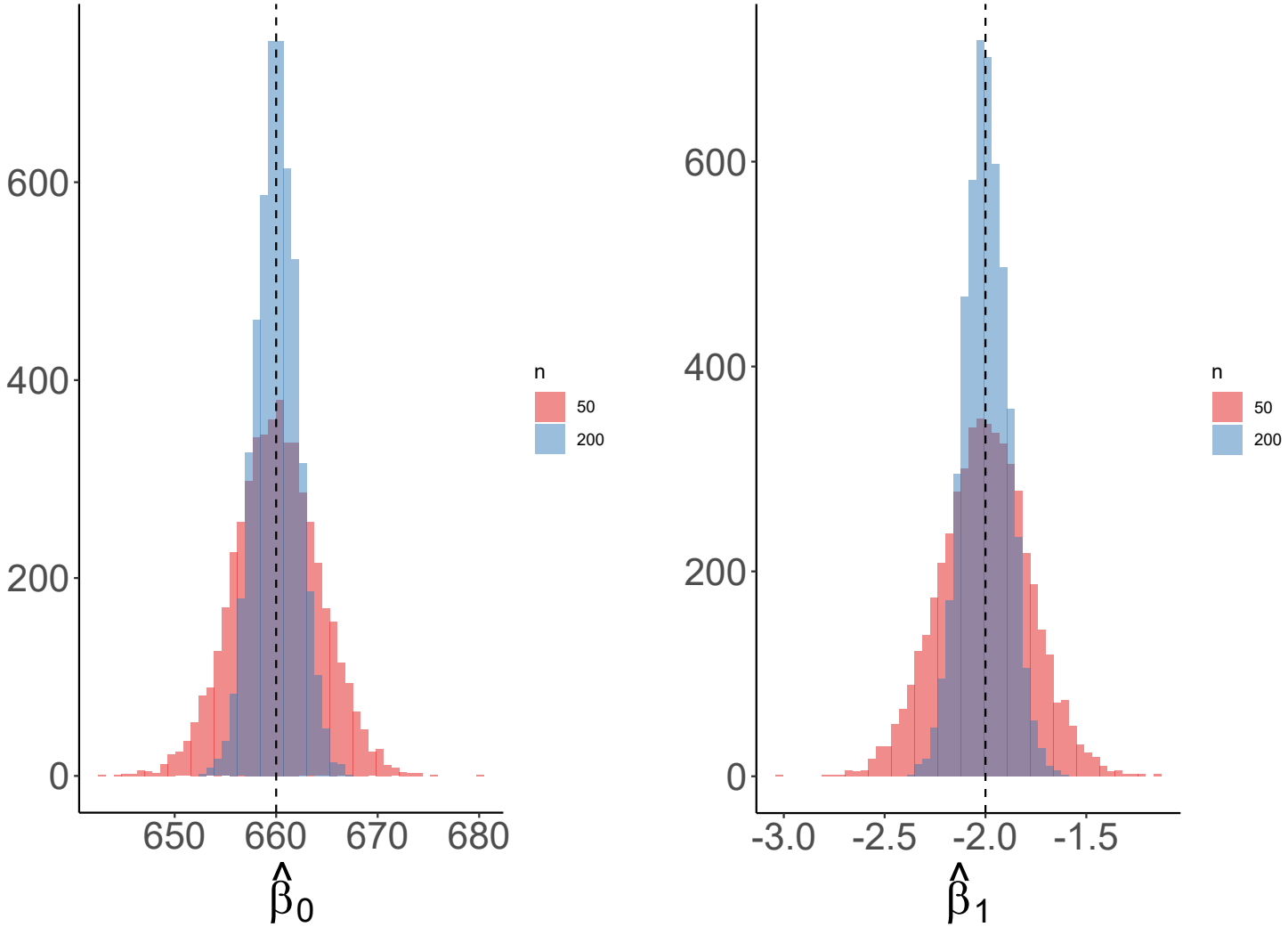
Симуляція Монте-Карло (6)

- Для того, щоб показати, що OLS оцінка є спроможна, збільшмо розмір вибірки з $n = 50$ до $n = 200$

```
df <- NULL
for (n in c(50, 200)){
  df <- rbind(df, as_tibble(t(replicate(T, get_beta(n, beta_0, beta_1, X = NULL)))))
}
df <- df %>% mutate(n = c(rep(50, T), rep(200, T)))
```

Симуляція Монте-Карло (7)

- Відповідні гістограми:



- Як можна бачити, оцінки справді спроможні
- Також цілком очевидно, що вони мають асимптотичний нормальний розподіл