Лекція 9. Регресійний аналіз - 3

Данило Тавров

05.04.2023

Вступні зауваги

- Сьогодні ми продовжуємо розглядати регресійний аналіз
- Корисними матеріалами є:
  - Фундаментальна книжка *Econometrics* (Bruce Hansen), розділи 2.4, 2.16, 2.17, (викладено на диску в загальному каталозі з літературою)
  - Книжка *Introduction to Econometrics* (James H. Stock, Mark W. Watson), розділи 8–9 (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, Данила Таврова та Ніколь Гандре

План лекції

Мотивація (1)

- Ми шукаємо причиново-наслідковий зв'язок між залежною змінною Y та однією або декількома незалежними змінними $\mathbf{X} = (1, X_1, \dots, X_k)^\top$
- Для цього ми розглядаємо функцію умовного сподівання (conditional expectation function, CEF)

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

  - Тут e — деяка похибка (інші невраховані фактори, які можуть впливати на Y, але які ми не врахували)
- На практиці як **апроксимацію** m ми розглядаємо лінійну модель

$$Y = \mathbf{X}^\top \beta + \tilde{e} \,, \qquad \mathbb{E}\left[\tilde{e} \mid \mathbf{X}\right] \approx 0$$

- Щоб **оцінити** β, ми шукаємо їх як коефіцієнти **лінійної проєкції** (або регресії) Y на $\mathbf{X}$:

$$Y = \mathbf{X}^\top \beta + u$$

- У результаті дістаємо коефіцієнти проєкції:

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

  - Принципово, що для лінійної проєкції виконується $\mathbb{E}\left[\mathbf{X}u\right] = 0$

Мотивація (2)

- 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 $\mathbf{X}_i^\top = (1, X_{1,i}, \dots, X_{k,i})$
- Тоді

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

Мотивація (3)

- Пригадаймо різницю між структурними моделями та лінійними проєкціями
- **Структурна** модель — це модель, яку ми хочемо оцінити:

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

  - У цій моделі **ми нічого не знаємо про** e!!!
  - **Якщо** $\mathbb{E}\left[e \mid \mathbf{X}\right] = 0$, то $\mathbb{E}\left[\mathbf{X}e\right] = 0$, і коефіцієнти *лінійної проєкції* будуть мати *причиново-наслідкову* інтерпретацію
- У **лінійній проєкції**

$$Y = \mathbf{X}^\top \gamma + u$$

  завжди $\mathbb{E}\left[\mathbf{X}u\right] = 0$
- Тобто якщо на практиці ми підозрюємо, що $\mathbb{E}\left[\mathbf{X}e\right] \neq 0$, то $\gamma \neq \beta$
- Ми говорили, що в цьому випадку з'являється *зміщення від неврахованої змінної* (omitted variable bias, OVB)
  - А регресор, який корелює з похибкою, називають **ендогенним** (endogenous)

Мотивація (4)

- Ми говорили, що для зменшення OVB в модель можна додати **контрольні змінні** (control variables):

$$Y_i = \mathbf{X}_i^\top \beta + \mathbf{W}_i^\top \delta + \tilde{e}_i$$

  - Ми вважаємо, що Cov $(\tilde{e}_i, \mathbf{X}_i) = 0$
- Контрольні змінні самі по собі інтересу для дослідника **не мають**
- Замість умови $\mathbb{E}[\tilde{e}_i \mid \mathbf{X}_i] = 0$ ми вимагаємо виконання $\mathbb{E}[\tilde{e}_i \mid \mathbf{X}_i, \mathbf{W}_i] = \mathbb{E}[\tilde{e}_i \mid \mathbf{W}_i]$
- У цьому випадку коефіцієнти β будуть мати причиново-наслідкову інтерпретацію для $\mathbf{X}_i$
- Можемо говорити, що, контролюючи $\mathbf{W}_i$, змінні X_i є «нічим не гірші від випадкових» (as good as randomly assigned)

Мотивація (5)

- Інший спосіб зменшити OVB та підвищити якість моделі загалом — додати **нелінійні ефекти** (nonlinear effects) змінних
- Тобто замість просто регресора X_i (наприклад, стать, рівень освіти, рівень доходу) можна додати деяку нелінійну функцію від X_i
  - Нас у першу чергу цікавитимуть поліноми, логаритми та добутки декількох змінних
- Модель $Y = \mathbf{X}^\top \beta + e$ від цього не перестане бути лінійною
  - Адже вона лінійна **в коефіцієнтах**, а не в регресорах
- Цей підхід трішки відрізняється від додавання контрольних змінних
  - І там, і там ми додаємо в модель нові змінні
  - Але додаючи нелінійні ефекти «основних» регресорів, ми намагаємося ліпше апроксимувати справжню CEF
  - Стверджувати, що ми аналізуємо вплив X, контролюючи (фіксуючи) значення X^2, було б просто безглуздо

Поліноми в лінійній регресії (1)

- Розгляньмо для простоти випадок, де нас цікавить вплив одного регресора X, і де ми розглядаємо вектор контрольних змінних $\mathbf{W}$:

$$Y = \beta_0 + \beta_1 X + \mathbf{W}^\top \delta + e$$

  - Ця модель за великим рахунком є структурною
  - Але для простоти вважатимемо, що контрольні змінні повністю нівелювали OVB, і тому $\mathbb{E}\left[e \mid X, \mathbf{W}\right] = \mathbb{E}\left[e \mid \mathbf{W}\right]$
- Тоді вплив X на Y дорівнює

$$\frac{\partial \mathbb{E}\left[Y \mid X, \mathbf{W}\right]}{\partial X} = \beta_1$$

  - Тобто від збільшення X на одну одиницю відбудеться збільшення (середнього) значення Y на β_1 одиниць
- Безпосереднім наслідком цих міркувань є те, що збільшення β_1 буде **однаковим**, незалежно від того, чому дорівнює X

Поліноми в лінійній регресії (2)

- Натомість, якщо додати в модель **ступені** X, матимемо

$$Y = \beta_0 + \beta_1 X + \beta_2 X^2 + ... + \beta_m X^m + \mathbf{W}^\top \delta + e$$

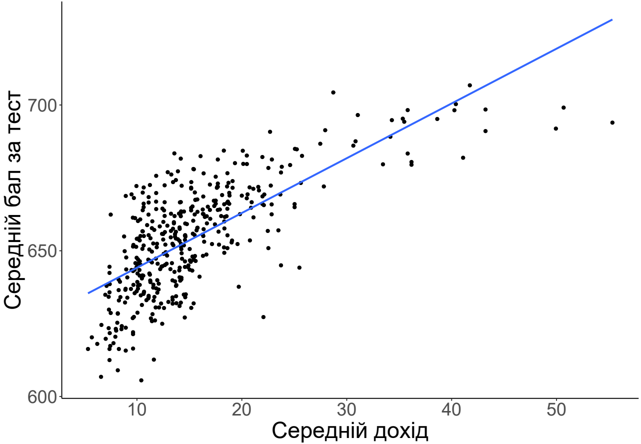
- І тоді вплив X на (середнє) значення Y дорівнюватиме

$$\frac{\partial \mathbb{E}\left[Y \mid X, \mathbf{W}\right]}{\partial X} = \beta_1 + 2\beta_2 X + ... + m\beta_m X^{m-1}$$

  - Тобто для різних значень X вплив буде різний

Приклад (1)

- Повернімося до датасету зі школами Каліфорнії
- Нехай нас цікавить залежність між середніми балами за випускний тест Y_i у шкільному окрузі та середніх доходом в окрузі на душу населення X_i
  - У датасеті змінна `avginc` відповідає середньому доході в окрузі в доларах 1998 р.
- Можемо збудувати залежність результатів тестів від середнього доходу



Приклад (2)

- Як можна бачити, що вищий середній достаток в окрузі, то вищі бали за тест
- Проте залежність не є лінійною
- Можна бачити, що збільшення доходу спочатку сильно пов'язано[1] зі збільшенням успішности
- Але десь на рівні 15 тис. доларів цей зв'язок починає слабшати
  - Фактично збільшення доходів більше не має значення, або навіть негативно пов'язано з балами за тести
- Відтак доречним видається моделювати таку залежність за допомогою квадратичної функції:

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

[1] Ми не кажемо **впливає**, бо це просто картинка з одним регресором, і робити причиново-наслідкову інтерпретацію було б явно передчасно!

Приклад (3)

- Можемо оцінити обидві моделі (лінійну й квадратичну)

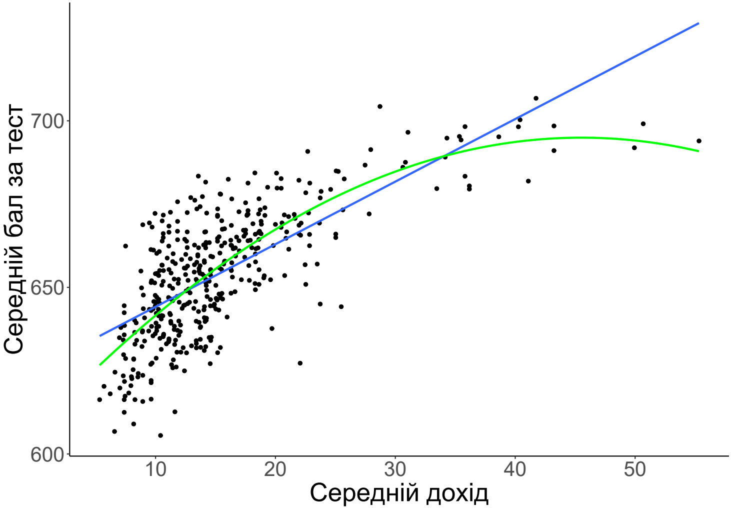
```
model <- lm(testscr ~ avginc, data = caschool)
model_hc1 <- coeftest(model, vcov. = hccm(model, type = "hc1"))

model2 <- lm(testscr ~ avginc + I(avginc^2), data = caschool)
model2_hc1 <- coeftest(model2, vcov. = hccm(model2, type = "hc1"))

stargazer(model, model2,
          type = "text",
          se = list(model_hc1[, 2], model2_hc1[, 2]),
          omit.stat = c("rsq", "f", "ser"),
          no.space = TRUE,
          digits = 3)
```

```
##
## =========================================
##                   Dependent variable:
##               ---------------------------
##                         testscr
##                    (1)            (2)
## -----------------------------------------
## avginc          1.879***      3.851***
##                 (0.114)        (0.268)
## I(avginc2)                    -0.042***
##                                (0.005)
## Constant       625.384***    607.302***
##                 (1.868)        (2.902)
## -----------------------------------------
## Observations      420            420
## Adjusted R2      0.506          0.554
## =========================================
## Note:         *p<0.1; **p<0.05; ***p<0.01
```

- Як можна бачити, коефіцієнт біля X^2 є статистично значущий
  - Це означає, що додавання X^2 до моделі було виправданим
  - Фактично, ми протестували гіпотезу $H_0 : \beta_2 = 0$ vs. $H_1 : \beta_2 \neq 0$ і відкинули її
  - Тобто лінійна модель **без** квадратичного терму поступається моделі **з** цим термом

Приклад (4)

- Також це можна побачити візуально



- Як можна бачити, квадратична модель ліпше описує наявні дані

Приклад (5)

- Як тепер варто інтерпретувати коефіцієнти моделі?
- Для лінійної моделі маємо $\widehat{\beta}_1 = 1.879$
  - Тобто збільшення середнього рівня доходів в окрузі на 1 тис. доларів **пов'язано** зі збільшенням середнього балу на 1.879
  - При цьому не важливо, чи в окрузі середній рівень доходу 5 тис. чи 55 тис.
- Для квадратичної моделі зв'язок дорівнює $\widehat{\beta}_1 + 2\widehat{\beta}_2 X = 3.851 - 0.085X$
  - Тобто збільшення з 10 тис. до 11 тис. пов'язано зі збільшенням середнього балу на 3.005
- Якщо бути зовсім коректним, то збільшення саме **на 1 тис. доларів —** це не зовсім значення похідної в точці
  - Насправді нам треба порахувати

$$\Delta\widehat{Y} = (\widehat{\beta}_0 + \widehat{\beta}_1 11 + \widehat{\beta}_2 11^2) - (\widehat{\beta}_0 + \widehat{\beta}_1 10 + \widehat{\beta}_2 10^2) = 2.963$$

- Аналогічні обчислення для ситуації, коли дохід збільшується з 40 до 41 тис. (використовуючи спрощену формулу через похідні), дають збільшення середнього бала за тест «всього» на 0.466

Приклад (6)

- Також корисним є оцінка «зв'язку» між X та Y для **середнього** (або медіанного) значення X
  - У нашому випадку ліпше взяти медіану, бо розподіли доходів завжди сильно скошені:

```
caschool %>% summarise(mean = mean(avginc), median = median(avginc))
```

```
## # A tibble: 1 x 2
##    mean median
##   <dbl>  <dbl>
## 1  15.3   13.7
```

- Для медіани X зв'язок дорівнює $\widehat{\beta}_1 + 2\widehat{\beta}_2 X = 2.689$
- Відтак, **якби** ці розрахунки мали причиново-наслідкову інтерпретацію, ми б могли зробити висновок, що для підвищення успішности учнів є сенс збільшувати рівень доходів у тих округах, де він не є занадто високий

Логарифми в лінійній регресії

- Інший поширений спосіб моделювання нелінійностей є використання (натуральних) логарифмів залежної та незалежної змінних
- Використання логарифмів дає змогу аналізувати **відносні** (відсоткові) залежності між змінними
- Типові приклади, де залежності саме такого роду становлять інтерес:
  - Аналіз різниці в доходах між різними категоріями працівників
  - Аналіз попиту та пропозиції на ринку
  - Інші види аналізу, де природними є відносні зміни, а не абсолютні

Лінійно-логарифмічна модель (1)

- Перший спосіб це зробити — збудувати **лінійно-логарифмічну** (linear-log) модель:
$$Y = \beta_0 + \beta_1 \ln X + \mathbf{W}^\top \delta + e$$
- Інтерпретувати коефіцієнт β_1 можна, використавши таку властивість логарифмів:
$$\ln(x + \Delta x) - \ln(x) = \ln\left(1 + \frac{\Delta x}{x}\right) \approx \frac{\Delta x}{x}$$
для малих Δx
    - Чисельно така апроксимація непогано працює для змін у межах 10%
    - Наприклад, $\ln(1.01) = 0.00995 \approx 0.01$ і $\ln(1.1) = 0.09531 \approx 0.1$
    - Але вже $\ln(1.5) = 0.40547 \neq 0.5$
- З урахуванням цього факту інтерпретація зв'язку така: збільшення X **на 1%** пов'язано зі збільшенням Y на $0.01\beta_1$ (ceteris paribus)

Лінійно-логаритмічна модель (2)

- Справді, розгляньмо ситуацію, коли X змінюється до $X + \Delta X$
- Тоді матимемо

$$\widehat{Y} = \widehat{\beta}_0 + \widehat{\beta}_1 \ln(X) + \mathbf{W}^\top \widehat{\delta}$$
$$\widehat{Y} + \Delta\widehat{Y} = \widehat{\beta}_0 + \widehat{\beta}_1 \ln(X + \Delta X) + \mathbf{W}^\top \widehat{\delta}$$

- Віднявши ці вирази, дістанемо

$$\Delta\widehat{Y} = \widehat{\beta}_1 \left(\ln(X + \Delta X) - \ln X\right) = \widehat{\beta}_1 \ln \frac{X + \Delta X}{X} \approx \widehat{\beta}_1 \frac{\Delta X}{X}$$

- Трішки неформально це ж можна побачити з виразу похідної для CEF:

$$\frac{\partial m}{\partial X} = \frac{\beta_1}{X} \quad \Rightarrow \quad \Delta m = \beta_1 \frac{\Delta X}{X}$$

- Можна помітити, що $\widehat{\beta}_1 \dfrac{\Delta X}{X} = 0.01\widehat{\beta}_1 \cdot 100\dfrac{\Delta X}{X}$
- Звідси $0.01\widehat{\beta}_1 \approx \dfrac{\Delta\widehat{Y}}{100\frac{\Delta X}{X}}$
- Оскільки $100\frac{\Delta X}{X}$ є ніщо інше, як **відносна** зміна X, виражена **у відсотках**, то маємо інтерпретацію, наведену на попередньому слайді

Приклад (1)

- У нашому прикладі зі школами замість додавання X^2 можемо використати лінійно-логаритмічну модель
- Можемо оцінити обидві моделі (лінійну й квадратичну)

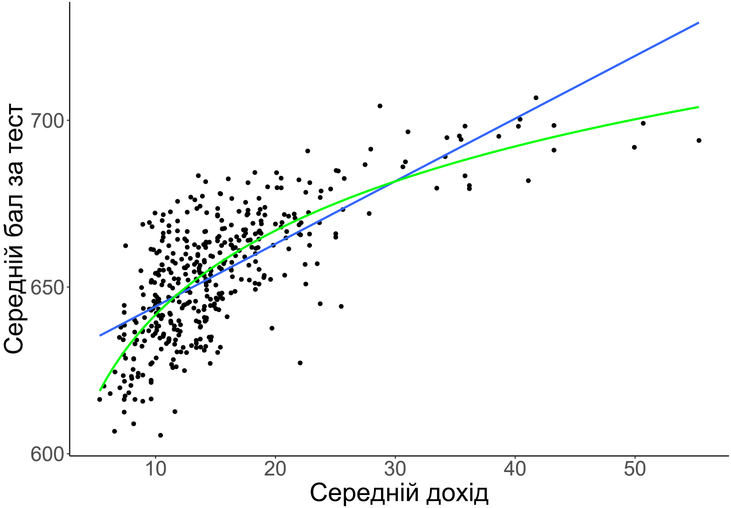
```
modellog <- lm(testscr ~ I(log(avginc)), data = caschool)
modellog_hc1 <- coeftest(modellog, vcov. = hccm(modellog, type = "hc1"))

stargazer(model, model2, modellog,
          type = "text",
          se = list(model_hc1[, 2], model2_hc1[, 2], modellog_hc1[, 2]),
          omit.stat = c("rsq", "f", "ser"),
          no.space = TRUE,
          digits = 3)
```

```
##
## ===============================================
##                       Dependent variable:
##                  ------------------------------
##                             testscr
##                     (1)        (2)        (3)
## -----------------------------------------------
## avginc           1.879***   3.851***
##                  (0.114)    (0.268)
## I(avginc2)                 -0.042***
##                             (0.005)
## I(log(avginc))                        36.420***
##                                        (1.397)
## Constant        625.384*** 607.302*** 557.832***
##                  (1.868)    (2.902)    (3.840)
## -----------------------------------------------
## Observations       420        420        420
## Adjusted R2       0.506      0.554      0.561
## ===============================================
## Note:               *p<0.1; **p<0.05; ***p<0.01
```

- Як можна бачити, коефіцієнт біля $\ln X$ є статистично значущий

Приклад (2)

- Також це можна побачити візуально



- Як можна бачити, лінійно-логаритмічна модель описує дані не гірше від квадратичної

Приклад (3)

- Як тепер варто інтерпретувати коефіцієнти моделі?
- Зв'язок між X та Y такий: збільшення середнього доходу на 1% пов'язано зі збільшенням середнього балу на 0.364
- Зокрема, збільшення з 10 до 11 тис. дорівнює збільшенню на $\frac{11-10}{10} = 10\%$, а тому пов'язано зі збільшенням балу на 3.642
  - Або, якщо бути зовсім коректним, то
$$\Delta\hat{Y} = (\hat{\beta}_0 + \hat{\beta}_1 \ln(10+1)) - (\hat{\beta}_0 + \hat{\beta}_1 \ln(10)) = 3.471$$
  - Квадратична модель нам давала зміну на 3.005
- Збільшення з 40 до 41 тис. є зміною на $\frac{41-40}{40} = 2.5\%$, а тому пов'язано зі збільшенням балу на 0.91
  - Квадратична модель нам давала зміну на 0.466
- У будь-якому випадку бачимо, що маржинальний приріст балів зменшується зі зростом середніх доходів

Логаритмічно-лінійна модель

- На відміну від лінійно-логаритмічної моделі, тепер ми беремо логаритм від залежної змінної:

$$\ln Y = \beta_0 + \beta_1 X + \mathbf{W}^\top \delta + e$$

- Інтерпретувати коефіцієнт β_1 можна в спосіб, схожий до попереднього
- Збільшення X **на 1 одиницю виміру** пов'язано зі збільшенням Y на $100\beta_1$% (ceteris paribus)
- Це можна вивести коректно за аналогією з попереднім випадком
- Матимемо

$$100\hat{\beta}_1 \approx \frac{100\frac{\Delta Y}{Y}}{\Delta X}$$

Приклад (1)

- Повернімося до датасету про зарплати працівників США

```
wages <- read_delim("data/cps09mar.csv", delim = ";") %>%
  mutate(hourly_wage = earnings / (hours*week)) %>%
  filter(hourly_wage >= 1) %>%
  mutate(log_hourly_wage = log(hourly_wage))
```

- Часто в контрактах пишуть, що зі збільшенням стажу працівник дістає **відсоткове** збільшення своєї зарплати
- Тоді логічним є взяти зарплату Y з логаритмом, а вік X — без

Приклад (2)

```
modelcps <- lm(hourly_wage ~ age, data = wages)
modelcps_hc1 <- coeftest(modelcps, vcov. = hccm(modelcps, type = "hc1"))

modelcps_log <- lm(log_hourly_wage ~ age, data = wages)
modelcps_log_hc1 <- coeftest(modelcps_log, vcov. = hccm(modelcps_log, type = "hc1"))

stargazer(modelcps, modelcps_log,
          type = "text",
          column.labels = c("Linear", "Log-linear"),
          se = list(modelcps_hc1[, 2], modelcps_log_hc1[, 2]),
          omit.stat = c("rsq", "f", "ser"),
          digits = 3)
```

```
##
## =========================================
##                  Dependent variable:
##              ----------------------------
##              hourly_wage  log_hourly_wage
##                 Linear       Log-linear
##                  (1)            (2)
## -----------------------------------------
## age            0.280***      0.011***
##                (0.007)       (0.0002)
##
## Constant      12.166***      2.498***
##                (0.292)        (0.011)
##
## -----------------------------------------
## Observations    50,628        50,628
## Adjusted R2     0.024          0.038
## =========================================
## Note:         *p<0.1; **p<0.05; ***p<0.01
```

Приклад (3)

- Як можна бачити, в обох моделях коефіцієнти статистично значущі
- Можна проаналізувати зв'язок в обох випадках
- Лінійна модель каже, що збільшення віку на 1 рік пов'язано зі збільшенням (середньої) погодинної зарплати на 0.28
  - Незалежно від поточного віку
- Логаритмічно-лінійна модель каже, що збільшення віку на 1 рік пов'язано зі збільшенням (середньої) погодинної зарплати на 1.088%
  - Незалежно від поточного віку

Логаритмічно-логаритмічна модель

- У цій моделі логаритмовано обидві змінні:

$$\ln Y = \beta_0 + \beta_1 \ln X + \mathbf{W}^\top \delta + e$$

- Інтерпретувати коефіцієнт β_1 можна у спосіб, схожий до попереднього
- Збільшення X **на 1 %** пов'язано зі збільшенням Y **на** β_1 **%** (ceteris paribus)
- Це можна вивести коректно за аналогією з попередніми випадками
- Матимемо

$$\hat{\beta}_1 \approx \frac{100\frac{\Delta Y}{Y}}{100\frac{\Delta X}{X}}$$

- Цей вираз називають **еластичністю** Y відносно X
  - Форми запису еластичності через похідні:

$$\frac{\partial Y}{\partial X} \cdot \frac{X}{Y} = \frac{\partial \ln Y}{\partial \ln X}$$

Приклад (1)

- Повернімося до нашого прикладу зі зв'язком доходів на бали за тести

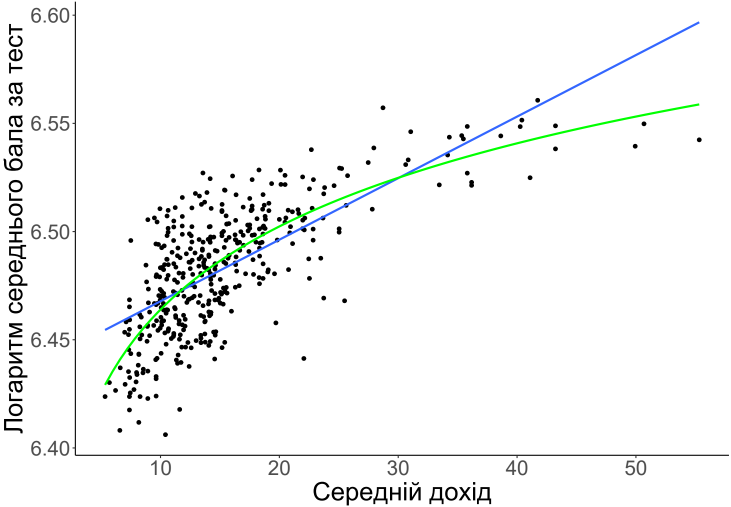
```r
modelloglin <- lm(I(log(testscr)) ~ avginc, data = caschool)
modelloglin_hc1 <- coeftest(modelloglin, vcov. = hccm(modelloglin, type = "hc1"))

modelloglog <- lm(I(log(testscr)) ~ I(log(avginc)), data = caschool)
modelloglog_hc1 <- coeftest(modelloglog, vcov. = hccm(modelloglog, type = "hc1"))

stargazer(model, modellog, modelloglog,
          type = "text",
          column.labels = c("Linear", "Linear-Log", "Log-Log"),
          se = list(model_hc1[, 2], modellog_hc1[, 2], modelloglog_hc1[, 2]),
          omit.stat = c("rsq", "f", "ser"),
          no.space = TRUE,
          digits = 3)
```

```
##
## =====================================================
##                          Dependent variable:
##                     ---------------------------------
##                         testscr        I(log(testscr))
##                     Linear   Linear-Log     Log-Log
##                      (1)        (2)           (3)
## -----------------------------------------------------
## avginc             1.879***
##                    (0.114)
## I(log(avginc))               36.420***     0.055***
##                               (1.397)       (0.002)
## Constant          625.384*** 557.832***    6.336***
##                    (1.868)    (3.840)       (0.006)
## -----------------------------------------------------
## Observations         420        420           420
## Adjusted R2         0.506      0.561         0.557
## =====================================================
## Note:                      *p<0.1; **p<0.05; ***p<0.01
```

- Як можна бачити, в усіх моделях коефіцієнти статистично значущі
- Збільшення сер. доходу на 1% пов'язано зі збільшенням сер. бала на 0.055%

Приклад (2)

- Також це можна побачити візуально



Приклад (3)

- Як можна бачити, логаритмічно-логаритмічна модель описує наявні дані ліпше від логаритмічно-лінійної
- Це ж можна побачити й за допомогою відповідних коефіцієнтів R^2_{adj}:

```
summary(modelloglin)$adj.r.squared
```

```
## [1] 0.4970107
```

```
summary(modelloglog)$adj.r.squared
```

```
## [1] 0.5567252
```

Загальні коментарі до логаритмічних моделей

- Логаритмічні моделі доцільно застосовувати тоді, коли цікавить зв'язок у відносному сенсі
- Також логаритмування рекомендовано, якщо розподіл змінної є сильно скошеним
  - Це підвищить якість лінійної апроксимації моделі
- Порівняння різних моделей між собою (наприклад, лінійну з лінійно-логаритмічною або логаритмічно-лінійну з логаритмічно-логаритмічною) можна здійснювати за допомогою R^2_{adj}
  - Але R^2_{adj} **не можна** застосовувати, якщо в одній моделі Y, а в іншій — $\ln Y$
- У загальному випадку вибір функційної форми моделі залежить від предметної області
  - Для зарплат доцільно брати $\ln Y$, бо відсоткова зміна зарплати цілком природна
  - Для тестових балів доцільно брати $\ln Y$, тому що не прийнято розглядати підвищення успішности учнів у відсоткових термінах

Порівняння поліномної та логаритмічної моделей на прикладі (1)

- Розгляньмо поліномну та логаритмічну моделі для нашого прикладу та порівняймо їх

```
model3 <- lm(testscr ~ avginc + I(avginc^2) + I(avginc^3), data = caschool)
model3_hc1 <- coeftest(model3, vcov. = hccm(model3, type = "hc1"))

modellinlog2 <- lm(testscr ~ I(log(avginc)) + I(log(avginc)^2), data = caschool)
modellinlog2_hc1 <- coeftest(modellinlog2, vcov. = hccm(modellinlog2, type = "hc1"))

modellinlog3 <- lm(testscr ~ I(log(avginc)) + I(log(avginc)^2) + I(log(avginc)^3), data = caschool)
modellinlog3_hc1 <- coeftest(modellinlog3, vcov. = hccm(modellinlog3, type = "hc1"))
```

Порівняння поліномної та логаритмічної моделей на прикладі (2)

```r
stargazer(model, model3, modellog, modellinlog2, modellinlog3,
          type = "text",
          column.labels = c("Linear", "Cubic", "Linear-Log", "Linear-Log-Quad", "Linear-Log-Cubic"),
          se = list(model_hc1[, 2], model3_hc1[, 2], modellog_hc1[, 2],
                    modellinlog2_hc1[, 2], modellinlog3_hc1[, 2]),
          omit.stat = c("rsq", "f", "ser"),
          no.space = TRUE,
          digits = 3)
```

```
##
## ==================================================================================
##                                         Dependent variable:
##                  -----------------------------------------------------------------
##                                                testscr
##                    Linear     Cubic    Linear-Log Linear-Log-Quad Linear-Log-Cubic
##                     (1)        (2)        (3)           (4)              (5)
## ----------------------------------------------------------------------------------
## avginc            1.879***   5.019***
##                   (0.114)    (0.707)
## I(avginc2)                  -0.096***
##                              (0.029)
## I(avginc3)                   0.001**
##                              (0.0003)
## I(log(avginc))                          36.420***     41.781***        113.381
##                                          (1.397)      (11.522)         (87.884)
## I(log(avginc)2)                                        -0.966          -26.911
##                                                        (2.011)         (31.746)
## I(log(avginc)3)                                                          3.063
##                                                                         (3.737)
## Constant         625.384*** 600.079*** 557.832***   550.562***       486.135***
##                   (1.868)    (5.102)    (3.840)      (16.260)         (79.383)
## ----------------------------------------------------------------------------------
## Observations        420        420        420           420              420
## Adjusted R2        0.506      0.555      0.561         0.561            0.560
## ==================================================================================
## Note:                                                *p<0.1; **p<0.05; ***p<0.01
```

Порівняння поліномної та логаритмічної моделей на прикладі (3)

- Як можна бачити, версії моделі з додаванням ступенів логаритмів не є корисними, бо всі коефіцієнти є статистично незначущими
  - Можна також виконати F-тест для всіх цих коефіцієнтів **одночасно**

```
linearHypothesis(modellinlog3, c("I(log(avginc)^2) = 0", "I(log(avginc)^3) = 0"),
                 vcov = hccm(modellinlog3, type = "hc1"))
```

```
## Linear hypothesis test
##
## Hypothesis:
## I(log(avginc)^2) = 0
## I(log(avginc)^3) = 0
##
## Model 1: restricted model
## Model 2: testscr ~ I(log(avginc)) + I(log(avginc)^2) + I(log(avginc)^3)
##
## Note: Coefficient covariance matrix supplied.
##
##   Res.Df Df      F Pr(>F)
## 1    418
## 2    416  2 0.4392 0.6448
```

  - Як видно, обидва коефіцієнти фактично є одночасно нульові, адже p-значення дуже високе
- Для поліномної моделі, натомість, навіть куб регресора має статистично значущий коефіцієнт
- Також можна порівняти відповідні R^2_{adj}
- Як видно, модель із логаритмами трішки ліпше описує дані
  - А оскільки вона містить тільки один регресор замість трьох, то її можна взяти як базову

Взаємодії між змінними (1)

- Коли ми обговорювали поліномні моделі, ми зазначали, що на в окремих ситуаціях зв'язок між X та Y не може бути сталим на всьому носії X
- У схожий спосіб можна заперечити, що в низці практичних ситуацій зв'язок X та Y може залежати від значень **інших** залежних змінних
  - Наприклад, ефект від зміни розміру класу може бути різний там, де багато дітей іммігрантів, і де їх мало
- Для того, щоб врахувати ефект від інших змінних, потрібно включати в модель **ефекти взаємодії** (interaction terms)
- Можна окремо розглянути три випадки:
  - Взаємодія між двома індикаторними змінними
  - Взаємодія між індикаторною та неперервною змінними
  - Взаємодія між двома неперервними змінними

Взаємодії між змінними: Дві індикаторні змінні (1)

- Розгляньмо для прикладу таку просту модель:

$$\ln W = \beta_0 + \beta_1 F + \beta_2 C + e$$

  - W — зарплата
  - F — жінка (1) чи чоловік (0)
  - C — вища освіта (1) чи ні (0)
- Інтерпретація коефіцієнтів OLS у такій моделі, як ми вже встановили вище, така:
  - $100\beta_1\%$ — різниця середньої зарплати між особами двох статей (контролюючи освіту)
  - $100\beta_2\%$ — різниця середньої зарплати між особами з вищою освітою та ні (контролюючи стать)
- Якщо ми вважаємо, що $\mathbb{E}\left[e \mid F, C\right] = 0$, то ми вважаємо, що змінної $F \times C$ у нашій моделі немає
- Трішки з іншого погляду, якщо ми *оцінюємо* таку модель через OLS, ми *насильно* задаємо значення 0 коефіцієнта біля $F \times C$
- Але чи виправдано це?

Взаємодії між змінними: Дві індикаторні змінні (2)

- Доволі сумнівно: вплив вищої освіти може бути різний для чоловіків і жінок
- Розгляньмо іншу модель:

$$\ln W = \beta_0 + \beta_1 F + \beta_2 C + \beta_3 F \times C + e$$

- Якщо $\mathbb{E}\left[e \mid F, C\right] = 0$, можемо говорити про вплив C на (умовне сподівання) $\ln W$, який дорівнює $\beta_2 + \beta_3 F$
  - Тобто для чоловіків вплив вищої освіти дорівнює $100\beta_2\%$
  - Для жінок вплив вищої освіти дорівнює $100(\beta_2 + \beta_3)\%$
- Аналогічно можемо говорити про вплив статі на зарплату залежно від рівня освіти

Взаємодії між змінними: Дві індикаторні змінні (3)

- Хоч у нашому датасеті про школи немає бінарних змінних, для ілюстрації можемо розглянути дві штучно утворені:
  - «Великий клас» `histr`, яка дорівнює 1, коли `str`≥ 20, та 0 — інакше
  - «Іммігрантський клас» `hiel`, яка дорівнює 1, коли `el_pct`$\geq 10\%$, та 0 — інакше
- Тоді маємо такі результати:

```
caschool <- caschool %>% mutate(histr = str >= 20, hiel = el_pct >= 10)

model_inter <- lm(testscr ~ histr + hiel, data = caschool)
model_inter_hc1 <- coeftest(model_inter, vcov. = hccm(model_inter, type = "hc1"))

model_inter_full <- lm(testscr ~ histr + hiel + histr:hiel, data = caschool)
# model_inter_full <- lm(testscr ~ histr*hiel, data = caschool)
model_inter_full_hc1 <- coeftest(model_inter_full, vcov. = hccm(model_inter_full, type = "hc1"))
```

Взаємодії між змінними: Дві індикаторні змінні (4)

```r
stargazer(model_inter, model_inter_full,
          type = "text",
          se = list(model_inter_hc1[, 2], model_inter_full_hc1[, 2]),
          omit.stat = c("rsq", "f", "ser"),
          digits = 3)
```

```
##
## ============================================
##                     Dependent variable:
##                 ----------------------------
##                           testscr
##                      (1)            (2)
## --------------------------------------------
## histr              -3.587**       -1.908
##                    (1.549)        (1.932)
##
## hiel              -19.719***    -18.163***
##                    (1.593)        (2.346)
##
## histrTRUE:hiel                    -3.494
##                                   (3.121)
##
## Constant          664.725***    664.143***
##                    (1.247)        (1.388)
##
## --------------------------------------------
## Observations         420            420
## Adjusted R2         0.290          0.290
## ============================================
## Note:            *p<0.1; **p<0.05; ***p<0.01
```

Взаємодії між змінними: Дві індикаторні змінні (5)

- Інтерпретація коефіцієнтів у другій моделі така:
  - Зв'язок між збільшенням розміру класу та тестовими балами дорівнює -1.908 для округів із незначним числом іммігрантів, і -5.402 — зі значним
- Щоправда, варто помітити, що в моделі з фактором взаємодії відповідний коефіцієнт є статистично незначущий
- Чи означає це, що потрібно його викинути і розглядати тільки першу модель?
- Необов'язково:
  - Якщо здоровий глузд чи теорія кажуть, що ефект повинен бути різний, то викидати нічого непотрібно
  - Коли ми вагалися, додавати квадрати й куби чи ні, ми намагалися підібрати точнішу апроксимацію CEF
  - Коли ми додаємо взаємодії між змінними, ми намагаємося здійснювати різні способи моделювання
  - Потрібно давати читачу самостійно зробити висновок, чи довіряти результатам, чи ні
- Ми про це ще поговоримо далі

Взаємодії між змінними: Індикаторна й неперервна змінні (1)

- Розгляньмо тепер випадок, коли один регресор бінарний, а інший — неперервний:

$$\ln W = \beta_0 + \beta_1 F + \beta_2 E + e$$

  - W — зарплата
  - F — жінка (1) чи чоловік (0)
  - E — рівень освіти (у роках)
- Інтерпретація коефіцієнтів OLS така:
  - $100\beta_1\%$ — різниця середньої зарплати між особами двох статей (контролюючи освіту)
  - $100\beta_2\%$ — різниця середньої зарплати від збільшення освіти на 1 рік (контролюючи стать)
- Розгляньмо модель із взаємодією:

$$\ln W = \beta_0 + \beta_1 F + \beta_2 E + \beta_3 F \times E + e$$

- Якщо $\mathbb{E}[e \mid F, E] = 0$, можемо говорити про вплив E на (умовне сподівання) $\ln W$, який дорівнює $\beta_2 + \beta_3 F$
  - Тобто для чоловіків вплив кожного року освіти дорівнює $100\beta_2\%$
  - Для жінок вплив вищої кожного року освіти дорівнює $100(\beta_2 + \beta_3)\%$
- Тобто якщо вдуматися, то для чоловіків і для жінок лінійна залежність (логарифмів) зарплати від освіти має різні кутові коефіцієнти

Взаємодії між змінними: Індикаторна й неперервна змінні (2)

- Розгляньмо для датасету зі школами регресори `str` та `hiel`
- Матимемо такі результати:

```
model_inter_bc <- lm(testscr ~ str + hiel, data = caschool)
model_inter_bc_hc1 <- coeftest(model_inter_bc, vcov. = hccm(model_inter_bc, type = "hc1"))

model_inter_bc_full <- lm(testscr ~ str + hiel + str:hiel, data = caschool)
# model_inter_bc_full <- lm(testscr ~ str*hiel, data = caschool)
model_inter_bc_full_hc1 <- coeftest(model_inter_bc_full, vcov. = hccm(model_inter_bc_full, type = "hc1

stargazer(model_inter_bc, model_inter_bc_full,
          type = "text",
          se = list(model_inter_bc_hc1[, 2], model_inter_bc_full_hc1[, 2]),
          omit.stat = c("rsq", "f", "ser"),
          no.space = TRUE,
          digits = 3)
```

```
##
## =========================================
##                   Dependent variable:
##              ----------------------------
##                         testscr
##                   (1)           (2)
## -----------------------------------------
## str            -1.491***      -0.968
##                 (0.475)       (0.589)
## hiel           -19.533***      5.639
##                 (1.558)      (19.515)
## str:hiel                      -1.277
##                               (0.967)
## Constant       692.361***   682.246***
##                 (9.557)      (11.868)
## -----------------------------------------
## Observations      420           420
## Adjusted R2      0.303         0.305
## =========================================
## Note:         *p<0.1; **p<0.05; ***p<0.01
```

Взаємодії між змінними: Індикаторна й неперервна змінні (3)

- Інтерпретація коефіцієнтів у другій моделі така:
  - Зв'язок між збільшенням розміру класу на 1 студента та тестовими балами дорівнює -0.968 для округів із незначним числом іммігрантів, і -2.245 — зі значним
- Варто помітити, що в моделі з фактором взаємодії відповідний коефіцієнт є статистично незначущий
- Проте додатково можемо протестувати декілька гіпотез
  - Чи є лінійні залежності балів від розміру класу для округів із великою та з малою кількістю мігрантів **абсолютно однакові**?
    - Тобто чи $\beta_{\texttt{hiel}} = \beta_{\texttt{str}\times\texttt{hiel}} = 0$?
  - Чи мають лінійні залежності балів від розміру класу **однакові кутові коефіцієнти**?
    - Тобто чи $\beta_{\texttt{str}\times\texttt{hiel}} = 0$?
  - Чи мають лінійні залежності балів від розміру класу **однакові перетини з віссю ординат**?
    - Тобто чи $\beta_{\texttt{hiel}} = 0$?

Взаємодії між змінними: Індикаторна й неперервна змінні (4)

- Перший тест дає такий результат:

```
linearHypothesis(model_inter_bc_full, c("hielTRUE = 0", "str:hielTRUE = 0"),
                 vcov = hccm(model_inter_bc_full, type = "hc1"))
```

```
## Linear hypothesis test
##
## Hypothesis:
## hielTRUE = 0
## str:hielTRUE = 0
##
## Model 1: restricted model
## Model 2: testscr ~ str + hiel + str:hiel
##
## Note: Coefficient covariance matrix supplied.
##
##   Res.Df Df      F    Pr(>F)
## 1    418
## 2    416  2 89.939 < 2.2e-16 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

- Ми **відкидаємо** цю гіпотезу, бо p-значення дуже мале
  - Тобто лінійні залежності **різні** для різних типів округів
- Проте окремі коефіцієнти статистично незначущі!
  - Тобто прямі мають однакові кутові коефіцієнти та перетини з віссю ординат

Взаємодії між змінними: Індикаторна й неперервна змінні (5)

- Чому так сталося?
- Насправді, ми маємо ситуацію з неповною мультиколінеарністю: кореляція між `hiel` та `str×hiel` дорівнює 0.993!
  - Відтак стандартні похибки цих коефіцієнтів є дуже великі
  - Тому важко сказати, **який саме** із них не дорівнює нулю
  - Але F-тест спільної гіпотези каже, що **принаймні** один із них точно ненульовий
- Загалом, **корисна порада**: тестуючи вплив деякого регресора, **потрібно виконувати спільний тест** з усіма коефіцієнтами, де цей регресор фігурує

Взаємодії між змінними: Дві неперервні змінні (1)

- Нарешті, якщо обидва регресори неперервні, маємо модель

$$\ln W = \beta_0 + \beta_1 X + \beta_2 E + \beta_3 X \times E + e$$

  - W — зарплата
  - X — досвід роботи (у роках)
  - E — рівень освіти (у роках)
- Інтерпретація коефіцієнтів OLS така:
  - $100(\beta_1 + \beta_3 E)\%$ — різниця середньої зарплати від збільшення досвіду робота на 1 рік (контролюючи освіту)
  - $100(\beta_2 + \beta_3 X)\%$ — різниця середньої зарплати від збільшення освіти на 1 рік (контролюючи досвід роботи)

Взаємодії між змінними: Індикаторна й неперервна змінні (2)

- Розгляньмо для датасету зі школами регресори `str` та `el_pct`
- Матимемо такі результати:

```r
model_inter_cont <- lm(testscr ~ str + el_pct, data = caschool)
model_inter_cont_hc1 <- coeftest(model_inter_cont, vcov. = hccm(model_inter_cont, type = "hc1"))

model_inter_cont_full <- lm(testscr ~ str + el_pct + str:el_pct, data = caschool)
# model_inter_cont_full <- lm(testscr ~ str*el_pct, data = caschool)
model_inter_cont_full_hc1 <- coeftest(model_inter_cont_full, vcov. = hccm(model_inter_cont_full, type

stargazer(model_inter_cont, model_inter_cont_full,
          type = "text",
          se = list(model_inter_cont_hc1[, 2], model_inter_cont_full_hc1[, 2]),
          omit.stat = c("rsq", "f", "ser"),
          no.space = TRUE,
          digits = 3)
```

```
##
## ==========================================
##                   Dependent variable:
##               ----------------------------
##                         testscr
##                    (1)            (2)
## ------------------------------------------
## str              -1.101**       -1.117*
##                  (0.433)        (0.588)
## el_pct          -0.650***       -0.673*
##                  (0.031)        (0.374)
## str:el_pct                       0.001
##                                 (0.019)
## Constant        686.032***    686.339***
##                  (8.728)       (11.759)
## ------------------------------------------
## Observations       420            420
## Adjusted R2       0.424          0.422
## ==========================================
## Note:          *p<0.1; **p<0.05; ***p<0.01
```

Взаємодії між змінними: Індикаторна й неперервна змінні (3)

- Інтерпретація коефіцієнтів у другій моделі така:
  - Зв'язок між збільшенням розміру класу на 1 студента та тестовими балами дорівнює, $-1.117 + 0.001 \cdot \texttt{el\_pct}$
  - Зокрема, для медіанного значення $\texttt{el\_pct}$, яке дорівнює 8.778, маємо -1.107
- І знову F-тест дає такий результат

```
linearHypothesis(model_inter_cont_full, c("el_pct = 0", "str:el_pct = 0"),
                 vcov = hccm(model_inter_cont_full, type = "hc1"))
```

```
## Linear hypothesis test
##
## Hypothesis:
## el_pct = 0
## str:el_pct = 0
##
## Model 1: restricted model
## Model 2: testscr ~ str + el_pct + str:el_pct
##
## Note: Coefficient covariance matrix supplied.
##
##   Res.Df Df      F    Pr(>F)
## 1    418
## 2    416  2 222.02 < 2.2e-16 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

- Тобто лінійні залежності **різні**, хоча **окремі** коефіцієнти статистично незначущі
- Також варто зазначити, що **розмір** впливу фактора взаємодії є доволі малий (усього 0.001)
  - Тобто навіть *якщо* цей коефіцієнт статистично значущий, особливого впливу він не має

Підсумкові зауваги

- Загальний підхід до використання нелінійних ефектів у регресійних моделях передбачає виконання таких кроків

1. З'ясувати, чи має місце (потенційний) нелінійний зв'язок між незалежною змінною та регресорами
   - Основним джерелом міркувань повинна бути теорія
   - Навіть до візуалізації потрібно проаналізувати, чи може існувати нелінійний зв'язок, яка його природа та форма
   - Візуалізація може тільки підтвердити початкові припущення
2. Додати в модель відповідні нелінійні ефекти та оцінити модель за допомогою OLS
3. Визначити, чи є нелінійна модель ліпшою від лінійної
   - Обов'язково потрібно емпірично підтверджувати будь-які теоретичні здогадки
   - Конкретніше, потрібно аналізувати, чи є статистично значущими коефіцієнт біля відповідних нелінійних ефектів
   - Якщо потрібно, варто застосувати F-тести для тестування, що *декілька* коефіцієнтів *сукупно* є статистично значущими
4. Здійснити візуалізацію, наклавши регресійну криву на діаграму розсіювання (scatter plot)
5. Оцінити зв'язок між зміною в регресорі X та зміною в залежній змінній
   - Як функцію від X та для деяких конкретних значень
   - Корисно брати середні або медіанні значення X

План лекції

Способи боротьби з OVB (1)

- Як ми вже вивчили, OVB виникатиме в ситуаціях, коли в структурній моделі $Y_i = \mathbf{X}_i\beta + e_i$ похибка e_i корелює з (принаймні одним із) $\mathbf{X}_i$
- Якщо в датасеті наявні дані про такі змінні, неврахованi в моделі, то їх потрібно включити в модель
- Інший варіант — додати контрольні змінні $\mathbf{W}_i$ такі, що для деякого фіксованого значення $\mathbf{W}_i$ змінні $\mathbf{X}_i$ є не гірші від випадково визначених
  - Тобто якщо $\mathbb{E}\left[e_i \mid \mathbf{X}_i, \mathbf{W}_i\right] = \mathbb{E}\left[e_i \mid \mathbf{W}_i\right]$
- Проте варто розуміти, що додавання змінних «про всяк випадок» є недоречним
- Що більше змінних включено в модель, то більша буде дисперсія OLS-оцінки
- Відтак, якщо деяка змінна насправді не повинна належати моделі, її не варто додавати
  - Це є приклад відомої проблеми балансування зміщення та дисперсії

Способи боротьби з OVB (2)

- На практиці доцільно слідувати таким чотирьом крокам

1. Визначити, які коефіцієнти становлять інтерес для дослідника
   - У нашому прикладі зі школами це є розмір класу
2. З'ясувати, чи існують джерела зміщення?
   - Для цього треба застосувати відому теорію
   - Або експертні знання
3. Додати нові змінні в базову модель, сформовану на кроці 1
   - Якщо коефіцієнти біля доданих контрольних змінних статистично значущі...
   - ...або якщо коефіцієнти біля основних змінних суттєво змінилися...
   - ...ці контрольні змінні потрібно залишити
   - Інакше їх включати не варто
4. Зобразити всі розрахунки у табличній формі, порівнюючи різні моделі між собою
   - Це дасть змогу читачу оцінити ситуацію та зробити власні висновки
   - Це є один із варіантів проведення так званих **перевірок на стійкість** (robustness checks)

Способи боротьби з OVB (3)

- Що робити, якщо змінних, які потрібно додати, у датасеті немає?

1. У датасеті може бути інформація про одну й ту саму одиницю спостереження в різні моменти часу
   - Такі дані називають **панельними** (panel)
   - Для роботи з панельними даними існують спеціальні методи, які розглядатимемо на наступній лекції
2. Можна знайти **інструментальні** змінні, які не пов'язані з основною моделлю, але дають змогу встановити причиново-наслідковий зв'язок
   - Ми про це говоритимемо наприкінці нашого курсу
3. Нарешті, можна провести власне дослідження RCT і забезпечити причиново-наслідкову інтерпретацію відповідної змінної

Способи боротьби з неправильною функційною формою

- Окрім OVB, неправильні результати можна дістати, якщо використовувати лінійну апроксимацію тоді, коли CEF насправді є суттєво нелінійною
  - Таку ситуацію називають **помилковою специфікацією функційної форми** (functional form misspecification)
- За великим рахунком, це також варіація на тему OVB
  - Але замість третіх змінних картину псує нелінійність у змінних, уже включених у модель
  - Наприклад, X^2 може корелювати з X, і якщо його не включити, матимемо неправильні оцінки коефіцієнтів
- Що з цим робити у випадку неперервних залежних змінних — ми розглянули вище
- Що з цим робити у випадку дискретних залежних змінних — розглядатимемо наступної лекції
  - Наприклад, якщо залежна змінна Y_i бінарна
  - У цьому випадку потрібно докласти зусиль, щоб наші прогнозні значення $\widehat{Y}_i$ щонайменше не виходили за межі інтервалу $[0; 1]$

Способи боротьби з похибками вимірювання (1)

- Інша причина, чому OLS-оцінки можуть бути викривлені, полягає в тому, що змінні можуть мати похибки вимірювання
  - Наприклад, респондент зазначає некоректну інформацію під час опитування
  - Або похибки може бути внесено на етапі занесення даних в електронну систему
- Розгляньмо, наскільки серйозною є така проблема
- Нехай маємо єдиний регресор X_i, замість якого ми спостерігаємо його викривлену версію $\tilde{X}_i$
- Тоді маємо таку модель:

$$Y_i = \beta_0 + \beta_1 \tilde{X}_i + v_i = \beta_0 + \beta_1 \tilde{X}_i + \left(\beta_1 \left(X_i - \tilde{X}_i\right) + u_i\right)$$

- Відтак, якщо $X_i - \tilde{X}_i$ корелює з $\tilde{X}_i$, матимемо OVB
  - Величина кореляції залежить від природи похибки вимірювання

Способи боротьби з похибками вимірювання (2)

- Наприклад, нехай $\tilde{X}_i = X_i + w_i$, $\mathbb{E}\left[w_i\right] = 0$, Var $\left(w_i\right) = \sigma_w^2$
- Також нехай Cov $\left(w_i, X_i\right) =$ Cov $\left(w_i, u_i\right) = 0$
- Тоді $v_i = \beta_1 \left(X_i - \tilde{X}_i\right) + u_i = -\beta_1 w_i + u_i$
- Також Cov $\left(\tilde{X}_i, w_i\right) =$ Cov $\left(X_i + w_i, w_i\right) = \sigma_w^2$
- Відтак Cov $\left(\tilde{X}_i, v_i\right) = -\beta_1$Cov $\left(\tilde{X}_i, w_i\right) +$ Cov $\left(\tilde{X}_i, u_i\right) = -\beta_1 \sigma_w^2$
- Підставляючи це в нашу формулу для OVB, дістаємо

$$\hat{\beta}_1 \xrightarrow{p} \beta_1 - \beta_1 \frac{\sigma_w^2}{\sigma_{\tilde{X}}^2} = \frac{\sigma_X^2}{\sigma_X^2 + \sigma_w^2} \beta_1$$

- Оскільки цей дріб менший від 1, наша оцінка $\hat{\beta}_1$ буде зміщена в бік 0

Способи боротьби з похибками вимірювання (3)

- Цікаво, що якщо похибка присутня в Y_i, тобто якщо $\tilde{Y}_i = Y_i + w_i$, то коефіцієнти будуть такі самі
  - Але дисперсія виросте
- Справді, маємо модель $\tilde{Y}_i = \beta_0 + \beta_1 X_i + v_i = \beta_0 + \beta_1 X_i + (w_i + u_i)$
- Якщо w_i незалежна ні від чого, то $\mathbb{E}\left[w_i \mid X_i\right] = 0$, а відтак $\mathbb{E}\left[v_i \mid X_i\right] = 0$, тобто OLS-оцінки будуть спроможними оцінками справжніх коефіцієнтів
- Але оскільки $\text{Var}\left(v_i\right) > \text{Var}\left(u_i\right)$, дисперсія $\hat{\beta}_1$ буде вищою

Способи боротьби з похибками вимірювання (4)

- У будь-якому випадку найліпший спосіб подолати цю проблему — дістати якнайточніші показники X_i
- Якщо це неможливо, то знову ж таки можуть стати в пригоді інструментальні змінні
- Також можна розробити математичну модель похибки (як це зроблено вище для найпростішого випадку)
  - Тоді, знаючи вираз для OVB, можна скоригувати оцінки
  - Наприклад, якщо в прикладі ми знаємо, що зміщення має коефіцієнт $\frac{\sigma_X^2}{\sigma_X^2+\sigma_w^2}$, то можна спробувати оцінити відповідні дисперсії
  - Тоді можна домножити на число, обернене до цього коефіцієнта, і спробувати скоригувати зміщення

Способи боротьби з одночасними причиново-наслідковими зв'язками (1)

- Ще однією неприємною проблемою в аналізі даних є ситуація, коли незрозуміло, що на що впливає
- Якщо X впливає на Y, і при цьому Y також впливає на X, кажуть, що мають місце **одночасні причиново-наслідкові зв'язки** (simultaneous causality)
- У цьому випадку коефіцієнти OLS не будуть спроможні
- У нашому випадку розмір класу, швидше за все, однозначно впливає на середні бал
- Проте можна уявити ситуацію, коли уряд визначає розмір класу залежно від рівня тестів
  - Наприклад, наймає більше вчителів в округи з поганими показниками
  - Тоді регресія в один бік буде некоректною

Способи боротьби з одночасними причиново-наслідковими зв'язками (2)

- Фактично, знову ж таки маємо OVB
- Нехай маємо дві (структурні!) моделі

$$Y_i = \beta_0 + \beta_1 X_i + u_i$$
$$X_i = \gamma_0 + \gamma_1 Y_i + v_i$$

- Тоді

$$\text{Cov}\left(X_i, u_i\right) = \text{Cov}\left(\gamma_0 + \gamma_1 Y_i + v_i, u_i\right) = \gamma_1 \text{Cov}\left(Y_i, u_i\right) + \text{Cov}\left(v_i, u_i\right)$$

- Навіть якщо $\text{Cov}\left(v_i, u_i\right) = 0$, маємо

$$\text{Cov}\left(X_i, u_i\right) = \gamma_1 \text{Cov}\left(\beta_0 + \beta_1 X_i + u_i, u_i\right) = \gamma_1 \beta_1 \text{Cov}\left(X_i, u_i\right) + \gamma_1 \sigma_u^2$$

- Звідси OVB дорівнює

$$\text{Cov}\left(X_i, u_i\right) = \frac{\gamma_1 \sigma_u^2}{1 - \gamma_1 \beta_1}$$

- Зокрема, воно дорівнює 0, якщо $\gamma_1 = 0$, тобто якщо Y_i справді не впливає на X_i

Способи боротьби з одночасними причиново-наслідковими зв'язками (3)

- Основних способів боротися з OVB такого типу є два
- Метод інструментальних змінних, який ми вивчатимемо наприкінці нашого курсу
- Організація повноцінного RCT, який повністю унеможливить причиново-наслідковий вплив в одному з напрямків

План лекції

Опис прикладу

- Розгляньмо пару прикладів з емпіричних вправ із підручника Stock & Watson
- Перший приклад базується на статті D. S. Hamermesh, A.Parker (2005). «Beauty in the classroom: instructors' pulchritude and putative pedagogical productivity», *Economics of Education Review*, 24(4), 369–376
    - Містить дані про оцінки викладачів за 463 дисциплін в Університеті Техасу в Остіні в 2000–2002 рр.
    - Датасет викачано звідси

```
evals <- read_csv("data/profs.csv")
```

```
## Rows: 463 Columns: 12
## -- Column specification --------------------------------------------------------
## Delimiter: ","
## chr (6): minority, gender, credits, division, native, tenure
## dbl (6): age, beauty, eval, students, allstudents, prof
##
## i Use `spec()` to retrieve the full column specification for this data.
## i Specify the column types or set `show_col_types = FALSE` to quiet this message.
```

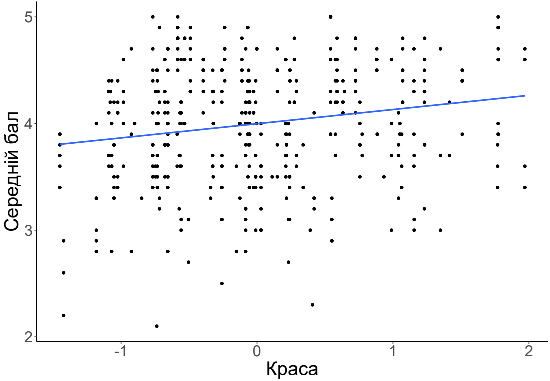
Опис датасету

- Змінні:
  - `minority`: чи є професор із нацменшини
  - `age`: вік професора
  - `gender`: стать професора (`female`, `male`)
  - `credits`: чи є дисципліна однокредитною вибірковою (`more`, `single`)
  - `beauty`: рейтинг привабливости викладача (середнє оцінок 6 студентів, центроване навколо 0)
  - `eval`: середній бал професора (від 1 до 5)
  - `division`: чи дисципліна читається на 1–2 курсах (`lower`), чи ні (`upper`)
  - `native`: чи є англійська рідною для професора (`no`, `yes`)
  - `tenure`: *дуже грубо кажучи*, чи працює професор в штаті (`yes`) чи ні (`no`)
  - `students`: число студентів, що взяли участь в оцінювання
  - `allstudents`: усього студентів, записаних на дисципліну
  - `prof`: ID професора
- Ми сконвертуємо окремі змінні в логічні

```
evals <- evals %>% mutate(female = ifelse(gender == "female", 1, 0),
                          one_credit = ifelse(credits == "single", 1, 0),
                          tenure = ifelse(tenure == "yes", 1, 0),
                          native = ifelse(native == "yes", 1, 0),
                          minority = ifelse(minority == "yes", 1, 0))
```

Візуалізація зв'язку

- Нас цікавить відповідь на таке дослідницьке питання: *Чи має вплив краса викладача на його оцінку?*
- Збудуймо відповідну діаграму розсіювання



- Як можна бачити, існує певний зв'язок
  - Наскільки він сильний?
  - Чи є він причиново-наслідковий?

Дескриптивні статистики

- Спочатку можемо подивитися на дескриптивні статистики

```
df <- as.data.frame(evals %>% dplyr::select(eval, beauty))
stargazer(df, type = "text",
          title = "Дескриптивні статистики", label = "table:sum-stats")
```

```
##
## Дескриптивні статистики
## ==========================================
## Statistic  N    Mean   St. Dev.  Min    Max
## ------------------------------------------
## eval      463  3.998   0.555   2.100  5.000
## beauty    463 0.00000  0.789   -1.450 1.970
## ------------------------------------------
```

- Використання `type = "latex"` замість `type = "text"` дає змогу згенерувати таблицю формату LaTeX, щоб результат виглядав професіональніше
  - Далі всі таблиці будуть подаватися в такому форматі

Таблиця 3.1: Summary Statistics

Statistic	N	Mean	St. Dev.	Min	Max
eval	463	3.998	0.555	2.100	5.000
beauty	463	0.00000	0.789	−1.450	1.970

Проста регресія (1)

- Розгляньмо відповідну регресію

```r
model <- lm(eval ~ beauty, data = evals)
model_hc1 <- coeftest(model, vcov. = hccm(model, type = "hc1"))

stargazer(model, type = "latex",
          title = "Проста регресія", label = "table:evals-reg",
          dep.var.labels = c("Середня оцінка"),
          dep.var.caption = "",
          se = list(model_hc1[, 2]),
          omit.stat = c("rsq", "f", "ser"))
```

Таблиця 3.2: Проста регресія

	Середня оцінка
beauty	0.133^{***}
	(0.032)
Constant	3.998^{***}
	(0.025)
Observations	463
R^2	0.036
Note:	$^{*}p<0.1$; $^{**}p<0.05$; $^{***}p<0.01$

Проста регресія (2)

- Можемо бачити, що збільшення бала за красу на 1 «пов'язано» зі збільшенням середньої оцінки викладача на 0.133
  - Цілком очевидно, що наявний OVB, тому поки що причиново-наслідкової інтерпретації в нас немає
- Чи можна вважати цей ефект великим?
- Щоб відповісти на це питання, можна проаналізувати відповідні середньоквадратичні відхилення:

```
beauty.sd <- sd(evals$beauty)
eval.sd <- sd(evals$eval)

model$coefficients[2] * beauty.sd / eval.sd
```

```
##    beauty
## 0.1890391
```

- Відтак збільшення бала за красу на 1 середньоквадратичне відхилення пов'язано зі збільшенням середньої оцінки на 18.904% від середньоквадратичного відхилення оцінки
  - Це доволі **мало**
  - Хоча коефіцієнт **статистично значущий**
- Чи можна казати, що бал за красу описує значну частку варіації в середній оцінці?
- Це показує коефіцієнт R^2, який у нашому випадку **дуже** малий
  - Тому **ні**

- Можемо збудувати довірчий інтервал навколо нашого коефіцієнта
- Це можна зробити як окремо:

```
ci <- coefci(model, vcov. = hccm(model, type = "hc1"))
ci
```

```
##                  2.5 %    97.5 %
## (Intercept) 3.94845765 4.0480866
## beauty      0.06949076 0.1965121
```

- Так і як частину великої таблиці:

```
stargazer(model, type = "text",
          title = "Проста регресія", label = "table:evals-reg2",
          dep.var.labels = c("Середня оцінка"),
          dep.var.caption = "",
          se = list(model_hc1[, 2]),
          omit.stat = c("rsq", "f", "ser"),
          ci = TRUE, ci.custom = list(ci),
          font.size = "tiny",
          no.space = TRUE)
```

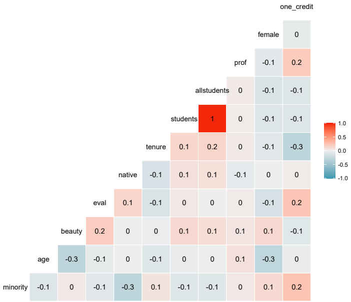
Таблиця 3.3: Проста регресія

	Середня оцінка
beauty	0.133^{***}
	(0.069, 0.197)
Constant	3.998^{***}
	(3.948, 4.048)
Observations	463
Adjusted R^2	0.034
Note:	*p<0.1; **p<0.05; ***p<0.01

Множинна регресія (1)

- Які контрольні змінні можна додати в нашу модель, щоб мінімізувати OVB?
- За великим рахунком, ніщо не заважає спробувати додати всі, що є
- Проте якщо трішки подумати, то стане зрозуміло, що не варто включати змінні, які сильно корелюють між собою

```
ggcorr(evals %>% dplyr::select(where(is.numeric)), label = TRUE)
```



- Як можна бачити, загальна кількість студентів і кількість студентів, що голосували, майже стовідсотково корелюють

Множинна регресія (2)

- Також дуже сумнівно, що той факт, що викладач працює в штаті чи ні, може впливати на його оцінку
  - Студенти часто цього не знають

```
model_mult <- lm(eval ~ beauty + minority + age + I(age^2) + female + one_credit + division + nat
model_mult_hc1 <- coeftest(model_mult, vcov. = hccm(model_mult, type = "hc1"))

model_mult_sig <- lm(eval ~ beauty + minority + female + one_credit + native, data = evals)
model_mult_sig_hc1 <- coeftest(model_mult_sig, vcov. = hccm(model_mult_sig, type = "hc1"))

stargazer(model, model_mult, model_mult_sig, type = "latex",
          title = "Множинна регресія", label = "table:evals-reg-mult",
          dep.var.labels = c("Середня оцінка"),
          dep.var.caption = "",
          se = list(model_hc1[, 2], model_mult_hc1[, 2], model_mult_sig_hc1[, 2]),
          omit.stat = c("rsq", "f", "ser"),
          no.space = TRUE,
          font.size = "tiny")
```

Множинна регресія (3)

Таблиця 3.4: Множинна регресія

	Середня оцінка		
	(1)	(2)	(3)
beauty	0.133^{***}	0.160^{***}	0.166^{***}
	(0.032)	(0.031)	(0.031)
minority		-0.180^{***}	-0.165^{**}
		(0.070)	(0.068)
age		0.020	
		(0.023)	
I(age^2)		-0.0002	
		(0.0002)	
female		-0.189^{***}	-0.174^{***}
		(0.052)	(0.050)
one_credit		0.617^{***}	0.641^{***}
		(0.111)	(0.097)
divisionupper		-0.004	
		(0.058)	
native		0.244^{**}	0.248^{***}
		(0.096)	(0.093)
students		-0.0001	
		(0.0005)	
Constant	3.998^{***}	3.436^{***}	3.824^{***}
	(0.025)	(0.559)	(0.095)
Observations	463	463	463
Adjusted R^2	0.034	0.141	0.145
Note:		$^{*}p<0.1$; $^{**}p<0.05$; $^{***}p<0.01$	

Множинна регресія (4)

- Як можна бачити, додавання контрольних змінних збільшило значення коефіцієнта біля бала за красу
- Хоча збільшення не є дуже суттєвим
- Різниця між специфікаціями (2) і (3) доволі несуттєва, тобто в даному випадку вилучення статистично незначущих регресорів було виправдано
- Між іншим, цікаво само по собі, що вік, кількість студентів та курс були статистично незначущі
- Вік потрібно протестувати явно

```
linearHypothesis(model_mult, c("age = 0", "I(age^2) = 0"), vcov. = hccm(model_mult, type = "hc1"))
```

```
## Linear hypothesis test
##
## Hypothesis:
## age = 0
## I(age^2) = 0
##
## Model 1: restricted model
## Model 2: eval ~ beauty + minority + age + I(age^2) + female + one_credit +
##     division + native + students
##
## Note: Coefficient covariance matrix supplied.
##
##   Res.Df Df      F Pr(>F)
## 1    455
## 2    453  2 0.6342 0.5308
```

- Як можна бачити, немає підстав відкинути гіпотезу, що вік **взагалі** не впливає

Нелінійні ефекти (1)

- Як можна модифікувати регресію, щоб ефект був різний для чоловіків і жінок?
- Для цього потрібно додати в модель взаємодію статі й бала за красу

```
model_inter <- lm(eval ~ beauty*female + minority + one_credit + native, data = evals)
model_inter_hc1 <- coeftest(model_inter, vcov. = hccm(model_inter, type = "hc1"))

stargazer(model, model_mult_sig, model_inter, type = "latex",
          title = "Множинна регресія з нелінійними ефектами", label = "table:evals-reg-female",
          dep.var.labels = c("Середня оцінка"),
          dep.var.caption = "",
          se = list(model_hc1[, 2], model_mult_sig_hc1[, 2], model_inter_hc1[, 2]),
          omit.stat = c("rsq", "f", "ser"),
          no.space = TRUE,
          font.size = "small")
```

Нелінійні ефекти (2)

Таблиця 3.5: Множинна регресія з нелінійними ефектами

	Середня оцінка		
	(1)	(2)	(3)
beauty	0.133^{***}	0.166^{***}	0.231^{***}
	(0.032)	(0.031)	(0.047)
minority		-0.165^{**}	-0.135^{*}
		(0.068)	(0.070)
female		-0.174^{***}	-0.173^{***}
		(0.050)	(0.049)
one_credit		0.641^{***}	0.656^{***}
		(0.097)	(0.096)
native		0.248^{***}	0.267^{***}
		(0.093)	(0.092)
beauty:female			-0.141^{**}
			(0.063)
Constant	3.998^{***}	3.824^{***}	3.807^{***}
	(0.025)	(0.095)	(0.094)
Observations	463	463	463
Adjusted R^2	0.034	0.145	0.153
Note:	$^{*}p<0.1$; $^{**}p<0.05$; $^{***}p<0.01$		

Нелінійні ефекти (3)

- Як можна бачити, ефект для двох статей суттєво відмінний, адже коефіцієнт біля взаємодії статистично значущий
- Можемо порахувати 95% довірчі інтервали для чоловіків та жінок окремо
- Для чоловіків усе просто: треба просто взяти інтервал для відповідного коефіцієнта:

```
ci <- coefci(model_inter, vcov. = hccm(model_inter, type = "hc1"))
ci["beauty", ]
```

```
##     2.5 %    97.5 %
## 0.1376500 0.3237619
```

- Для жінок потрібно знайти інтервал для $\hat{\beta}_{\texttt{beauty}} + \hat{\beta}_{\texttt{beauty:female}}$
  - Для цього треба знати відповідну стандартну похибку
  - Її нескладно порахувати, адже ми знаємо повністю всю матрицю коваріацій для коефіцієнтів $\hat{\beta}$
  - Звідти можна дістати відповідні дисперсії та коваріації і порахувати $\text{se}\left(\hat{\beta}_{\texttt{beauty}} + \hat{\beta}_{\texttt{beauty:female}}\right) = \sqrt{\text{Var}\left(\hat{\beta}_{\texttt{beauty}}\right) + \text{Var}\left(\hat{\beta}_{\texttt{beauty:female}}\right) + 2\text{Cov}\left(\hat{\beta}_{\texttt{beauty}}, \hat{\beta}_{\texttt{beauty:female}}\right)}$

Нелінійні ефекти (4)

- Автоматично це можна зробити за допомогою функції `glht` з пакета `multcomp`:

```
lin.test <- glht(model_inter, c("beauty + beauty:female = 0"),
                 vcov. = hccm(model_inter, type = "hc1"))
confint(lin.test)
```

```
##
##   Simultaneous Confidence Intervals
##
## Fit: lm(formula = eval ~ beauty * female + minority + one_credit +
##     native, data = evals)
##
## Quantile = 1.9652
## 95% family-wise confidence level
##
##
## Linear Hypotheses:
##                             Estimate lwr     upr
## beauty + beauty:female == 0 0.09011  0.01145 0.16877
```

- Як можна бачити, 0 не належить цьому інтервалу, отже для жінок вплив також статистично значущий
  - Хоча й невеликий