Лекція 8. Регресійний аналіз - 2

Данило Тавров

29.03.2023

Вступні зауваги

- Сьогодні ми продовжуємо розглядати регресійний аналіз
- Корисними матеріалами є:
  - Фундаментальна книжка *Econometrics* (Bruce Hansen), розділи 7, 9 (викладено на диску в загальному каталозі з літературою)
  - Книжка *Introduction to Econometrics* (James H. Stock, Mark W. Watson), розділи 5, 7, 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 та Данила Таврова

План лекції

1. Зміщення від неврахованих змінних
2. Статистичне виведення для коефіцієнтів регресії

Різниця між структурними моделями та лінійними проєкціями (1)

- Пригадаймо, чим ми займаємося в рамках регресійного аналізу даних
- Нас цікавить встановити причиново-наслідковий зв'язок між змінними
  - Ми можемо встановити такий зв'язок у деякому «середньому» сенсі
  - Для цього ми розглядаємо функцію умовного сподівання (conditional expectation function, CEF)
- Нехай ми маємо модель

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

  - Тут Y — залежна змінна, вплив на яку ми хочемо проаналізувати
  - $\mathbf{X} = (X_1, \dots, X_k)^\top$ — незалежні змінні
  - e — деяка похибка (інші невраховані фактори, які можуть впливати на Y, але які ми не врахували)
- **Якщо** виконується умова про нульове умовне сподівання, то $m(\mathbf{X}) = \mathbb{E}\left[Y \mid \mathbf{X}\right]$
  - Зверніть увагу, що ми не вимагаємо **незалежности** e від $\mathbf{X}$
  - Це було б занадто
  - Згадайте, що похибки гетероскедастичні, коли Var $(e \mid \mathbf{X})$ не є сталою
- Тоді можна взяти похідну $\frac{\partial m}{\partial X_i}$ та інтерпретувати її як вплив змінної X_i
  - Оскільки це частинна похідна, усі інші змінні є фіксовані
  - І відповідна похідна має інтерпретацію впливу *ceteris paribus* («за інших рівних умов»)
- За законом ітерованих сподівань, $\mathbb{E}\left[e\mathbf{X}\right] = \mathbb{E}\left[\mathbb{E}\left[\mathbf{X}e \mid \mathbf{X}\right]\right] = \mathbb{E}\left[\mathbf{X}\mathbb{E}\left[e \mid \mathbf{X}\right]\right] = 0$
  - Тобто якщо $\mathbb{E}\left[e\mathbf{X}\right] \neq 0$, то умова про нульове умовне сподівання **не** виконується
  - Іншими словами, $m(\mathbf{X})$ **не буде** CEF, якщо e є **корельована** з $\mathbf{X}$

Різниця між структурними моделями та лінійними проєкціями (3)

- Ми показували, що суто алгебрично виходить, що $\mathbb{E}\left[\mathbf{X}u\right] = 0$
  - Тобто похибка проєкції u завжди некорельована з регресорами $\mathbf{X}$
- Оцінкою коефіцієнтів проєкції є відповідна plug-in оцінка
  - Вона ж — оцінка найменших квадратів (ordinary least squares estimator, OLS)
- Нехай маємо вибірку $(Y_i, X_{1,i}, \ldots, X_{k,i})$, $i = 1, \ldots, n$
  - Вважаємо, що всі n векторів незалежні **між собою**
  - ...та мають однаковий розподіл $\mathbb{P}_{Y, X_1, \ldots, X_k}$, який нам невідомий
- Нехай $\mathbf{X}_i^\top = (1, X_{1,i}, \ldots, X_{k,i})$
- Тоді

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

Різниця між структурними моделями та лінійними проєкціями (4)

- Потрібно дуже чітко розрізняти структурні моделі та всього лише лінійні проєкції
- Нехай нас цікавить оцінити вплив статі X_1, раси X_2 та освіти X_3 на зарплату Y
- Ми воліємо розглянути лінійну модель відповідної CEF:

$$Y = \beta_0 + \beta_1 X_1 + \beta_2 X_2 + \beta_3 X_3 + e$$

- У цій моделі **ми нічого не знаємо про** e!!!
  - Ця модель є **структурна**, ми не заявляємо, що ми записали справжню CEF
  - Якби було $\mathbb{E}\left[e \mid X_1, X_2, X_3\right] = 0$, то це було б ідеально
  - Але існують такі змінні, як вроджені здібності чи галузь діяльности, що **корелюють** з X_1, X_2, X_3
  - Відтак $\mathbb{E}\left[e \mid X_1, X_2, X_3\right] \neq 0$
  - Це не вирок!
  - Це виклик, із яким потрібно боротися, і для цього існують різні методи, деякі з яких ми будемо вивчати
- Але **оцінювання** коефіцієнтів цієї моделі ми робимо за методом OLS, тобто ми **припускаємо**, що

$$Y = \gamma_0 + \gamma_1 X_1 + \gamma_2 X_2 + \gamma_3 X_3 + u \;, \qquad \mathbb{E}\left[\mathbf{X} u\right] = 0$$

  - Якщо умова про нульову кореляцію u з деяким X_j не виконується, то $\beta_j \neq \gamma_j$!

Зміщення від неврахованої змінної (1)

- Цю різницю можна квантифікувати
- Нехай маємо модель з одним регресором: $Y_i = \beta_0 + \beta_1 X_i + e_i$
- Нехай e_i містить у собі деяку іншу змінну W_i таку, що Cov $(X_i, e_i) \neq 0$
- Розгляньмо лінійну проєкцію $Y_i = \gamma_0 + \gamma_1 X_i + u_i$, $\mathbb{E}[X_i u_i] = 0$
- Тоді можна записати

$$\gamma_1 = \frac{\text{Cov}(Y_i, X_i)}{\text{Var}(X_i)} = \frac{\text{Cov}(\beta_0 + \beta_1 X_i + e_i, X_i)}{\text{Var}(X_i)}$$
$$= \frac{\beta_1 \text{Var}(X_i) + \text{Cov}(e_i, X_i)}{\text{Var}(X_i)} \neq \beta_1$$

- Зміщення $\frac{\text{Cov}(X_i, e_i)}{\text{Var}(X_i)}$ називають **зміщенням від неврахованої змінної** (omitted variable bias, OVB)
- Регресор X_i називають **ендогенним** (endogenous)
  - Тому що його значення не є повністю незалежним, а частково визначається самою моделлю
  - Проблема **ендогенности** (endogeneity) — найважливіша проблема в регресійному аналізі[1]

[1] Та економетриці загалом

Зміщення від неврахованої змінної (2)

- Можемо проаналізувати знак такого зміщення
- Нехай у нашій структурній моделі $e_i = \beta_2 W_i + u_i$
    - Вважаємо Cov $(X_i, u_i) = 0$, Cov $(W_i, u_i) = 0$
    - Тобто повна модель була б $Y_i = \beta_0 + \beta_1 X_i + \beta_2 W_i + u_i$
    - І її коефіцієнти можна оцінити за допомогою OLS (вони будуть спроможні)
- Тоді зміщення дорівнює $\frac{\text{Cov}(X_i, e_i)}{\text{Var}(X_i)} = \beta_2 \frac{\text{Cov}(X_i, W_i)}{\text{Var}(X_i)}$
- Оскільки дисперсія завжди додатна, знак OVB визначають:
    - Знак Cov (X_i, W_i), тобто додатна чи від'ємна кореляція між неврахованою змінною W_i і змінною X_i
    - Знак β_2, тобто додатна чи від'ємна кореляція між неврахованою змінною W_i і залежною змінною Y_i

Зміщення від неврахованої змінної (3)

- Нехай маємо модель із багатьма регресорами: $Y_i = \mathbf{X}_i^\top \beta + \mathbf{W}_i^\top \delta + e_i$
  - Тут $\mathbf{X}_i^\top = (1, X_{1,i}, \dots, X_{k,i})$
  - $\mathbf{W}_i^\top = (W_{1,i}, \dots, W_{p,i})$
  - $\mathbb{E}\left[e_i \mid \mathbf{X}_i, \mathbf{W}_i\right] = 0$
  - Цю регресію інколи називають **довгою регресією** (long regression)
- Нехай замість цієї моделі ми оцінюємо модель $Y_i = \mathbf{X}_i^\top \gamma + u_i$, $\mathbb{E}\left[\mathbf{X}_i u_i\right] = 0$
  - Цю регресію інколи називають **короткою регресією** (short regression)
- Тоді можемо обчислити коефіцієнти γ за звичною формулою:

$$
\begin{aligned}
\gamma &= \left(\mathbb{E}\left[\mathbf{X}_i \mathbf{X}_i^\top\right]\right)^{-1} \mathbb{E}\left[\mathbf{X}_i Y_i\right] = \left(\mathbb{E}\left[\mathbf{X}_i \mathbf{X}_i^\top\right]\right)^{-1} \mathbb{E}\left[\mathbf{X}_i \left(\mathbf{X}_i^\top \beta + \mathbf{W}_i^\top \delta + e_i\right)\right] \\
&= \beta + \left(\mathbb{E}\left[\mathbf{X}_i \mathbf{X}_i^\top\right]\right)^{-1} \mathbb{E}\left[\mathbf{X}_i \mathbf{W}_i^\top\right] \delta = \beta + \Gamma \delta
\end{aligned}
$$

- Γ — це матриця розмірности $(k+1) \times p$
- Можна помітити, що коефіцієнти Γ є коефіцієнтами проєкції $\mathbf{W}_i$ на $\mathbf{X}_i$
- Ми матимемо OVB $\Gamma\delta = 0$ у двох випадках
  - Або $\Gamma = 0$, тобто $\mathbf{W}_i$ і $\mathbf{X}_i$ некорельовані
  - Або $\delta = 0$, тобто $\mathbf{W}_i$ насправді в нашу модель і не повинні входити, бо не впливають на Y_i
- В усіх інших випадках матимемо ненульовий OVB

Приклад (1)

- Розгляньмо приклад із попередньої лекції про результати тестування з читання та математики учнів 5-х класів шкіл Каліфорнії (1999 р.)[2]

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

```
## Rows: 420 Columns: 18
## -- Column specification --------------------------------------------------------
## Delimiter: ","
## chr  (3): county, district, gr_span
## dbl (15): observation_number, dist_cod, enrl_tot, teachers, calw_pct, meal_p...
##
## 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.
```

[2] Приклад узято з книжки Stock & Watson

Приклад (2)

- Ми аналізували структурну модель $Y_i = \beta_0 + \beta_1 X_i + e_i$
  - 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`
- Оцінки цих коефіцієнтів за OLS дорівнюють

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

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

- Проте нас беруть справедливі сумніви, що навряд чи похибка e_i некорельована з X_i
  - Ми не врахували таких факторів, як якість учителів, матеріально-технічне забезпечення, соціальний стан дітей тощо
- Розгляньмо для прикладу одну змінну, притаманну цьому датасету
  - У школах Каліфорнії часто можуть учитися діти мігрантів із Мексики (особливо нелегальних)
  - Ці діти часто навіть не володіють англійською мовою
  - Відповідно, їхні результати за тести будуть нижчі

Приклад (3)

- Така змінна в нашому датасеті — `el_pct`
- Можемо порахувати відповідні коваріації та дисперсію:

```
cov(caschool$el_pct, caschool$str)
```

```
## [1] 6.491213
```

```
var(caschool$str)
```

```
## [1] 3.578952
```

```
cov(caschool$el_pct, caschool$str) / var(caschool$str)
```

```
## [1] 1.813719
```

- Коваріація додатна
  - Можемо очікувати, що $\beta_2 < 0$
  - Тобто вплив значного числа неангломовних студентів на тести від'ємний
  - Відтак OVB повинно бути від'ємне
- Справді,

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

```
##
## Call:
## lm(formula = testscr ~ str + el_pct, data = caschool)
##
## Coefficients:
## (Intercept)          str       el_pct
##    686.0322      -1.1013      -0.6498
```

- Як можна бачити, після врахування змінної `el_pct` вплив розміру класу став удвічі менший!

Ілюстрація методом Монте-Карло (1)

- Для ілюстрації впливу OVB розгляньмо симуляцію Монте-Карло для такого штучного прикладу
- Нехай маємо $X_i \sim \text{Bern}\,(0.7)$, $W_i \sim \text{Bern}\,(0.4)$, $X_i \perp\!\!\!\perp W_i$
- Розгляньмо дві різні моделі:

$$Y_{1,i} = 0.2 + 0.6X_i + 0.7W_i + e_i\ , \qquad \mathbb{E}\left[e_i \mid X_i, W_i\right] = 0$$
$$Y_{2,i} = 0.2 + 0.6X_i + 0.7W_i - 0.3X_iW_i + u_i\ , \qquad \mathbb{E}\left[u_i \mid X_i, W_i\right] = 0$$

  - Тобто обидві ці моделі є повністю істинними
- У рамках нашої симуляції будемо $T = 5000$ разів генерувати відповідні вибірки та оцінювати коефіцієнти
  - Для обох моделей оцінюватимемо коефіцієнти OLS, використовуючи тільки X_i і W_i, без X_iW_i
  - У першому випадку оцінки будуть спроможні
  - У другому — ні

Ілюстрація методом Монте-Карло (2)

- Можемо обчислити відповідні OVB
- Справді,

$$
\begin{aligned}
\Gamma &= \left(\mathbb{E}\left[\begin{pmatrix}1\\X_i\\W_i\end{pmatrix}(1, X_i, W_i)\right]\right)^{-1}\mathbb{E}\left[\begin{pmatrix}1\\X_i\\W_i\end{pmatrix}X_iW_i\right]\\
&= \begin{pmatrix}1 & \mathbb{E}[X_i] & \mathbb{E}[W_i]\\ \mathbb{E}[X_i] & \mathbb{E}\left[X_i^2\right] & \mathbb{E}[X_iW_i]\\ \mathbb{E}[W_i] & \mathbb{E}[W_iX_i] & \mathbb{E}\left[W_i^2\right]\end{pmatrix}^{-1}\begin{pmatrix}\mathbb{E}[X_iW_i]\\ \mathbb{E}\left[X_i^2W_i\right]\\ \mathbb{E}\left[X_iW_i^2\right]\end{pmatrix}
\end{aligned}
$$

- Тут $\mathbb{E}\left[X_i^2\right] = 1^2 \cdot \mathbb{P}_X(X_i = 1) + 0 \cdot \mathbb{P}_X(X_i = 0) = 0.7$
- $\mathbb{E}\left[W_i^2\right] = 1^2 \cdot \mathbb{P}_W(W_i = 1) + 0 \cdot \mathbb{P}_W(W_i = 0) = 0.4$
- За формулою повного сподівання,

$$
\begin{aligned}
\mathbb{E}[X_iW_i] &= \mathbb{E}[X_iW_i \mid W_i = 1]\,\mathbb{P}_W(W_i = 1) + \mathbb{E}[X_iW_i \mid W_i = 0]\,\mathbb{P}_W(W_i = 0)\\
&= \mathbb{E}[X_i]\,\mathbb{P}_W(W_i = 1) = 0.7 \cdot 0.4 = 0.28
\end{aligned}
$$

- Аналогічно $\mathbb{E}\left[X_i^2W_i\right] = \mathbb{E}\left[X_iW_i^2\right] = 0.28$

Ілюстрація методом Монте-Карло (2)

- Можемо обчислити відповідні OVB
- Справді,

$$\Gamma = \left(\mathbb{E}\left[\begin{pmatrix}1\\X_i\\W_i\end{pmatrix}(1, X_i, W_i)\right]\right)^{-1}\mathbb{E}\left[\begin{pmatrix}1\\X_i\\W_i\end{pmatrix}X_iW_i\right]$$
$$= \begin{pmatrix}1 & \mathbb{E}[X_i] & \mathbb{E}[W_i]\\ \mathbb{E}[X_i] & \mathbb{E}\left[X_i^2\right] & \mathbb{E}[X_iW_i]\\ \mathbb{E}[W_i] & \mathbb{E}[W_iX_i] & \mathbb{E}\left[W_i^2\right]\end{pmatrix}^{-1}\begin{pmatrix}\mathbb{E}[X_iW_i]\\ \mathbb{E}\left[X_i^2W_i\right]\\ \mathbb{E}\left[X_iW_i^2\right]\end{pmatrix}$$

  - Тут $\mathbb{E}\left[X_i^2\right] = 1^2 \cdot \mathbb{P}_X(X_i = 1) + 0 \cdot \mathbb{P}_X(X_i = 0) = 0.7$
  - $\mathbb{E}\left[W_i^2\right] = 1^2 \cdot \mathbb{P}_W(W_i = 1) + 0 \cdot \mathbb{P}_W(W_i = 0) = 0.4$
  - За формулою повного сподівання,

$$\mathbb{E}[X_iW_i] = \mathbb{E}[X_iW_i \mid W_i = 1]\,\mathbb{P}_W(W_i = 1) + \mathbb{E}[X_iW_i \mid W_i = 0]\,\mathbb{P}_W(W_i = 0)$$
$$= \mathbb{E}[X_i]\,\mathbb{P}_W(W_i = 1) = 0.7 \cdot 0.4 = 0.28$$

  - Аналогічно $\mathbb{E}\left[X_i^2W_i\right] = \mathbb{E}\left[X_iW_i^2\right] = 0.28$

Ілюстрація методом Монте-Карло (3)

- Отже OVB дорівнює

$$\Gamma\delta = \begin{pmatrix} 1 & 0.7 & 0.4 \\ 0.7 & 0.7 & 0.28 \\ 0.4 & 0.28 & 0.4 \end{pmatrix}^{-1} \begin{pmatrix} 0.28 \\ 0.28 \\ 0.28 \end{pmatrix} \cdot (-0.3) = \begin{pmatrix} 0.084 \\ -0.12 \\ -0.21 \end{pmatrix}$$

Ілюстрація методом Монте-Карло (4)

- Функція для генерації датасетів

```
generate_data <- function(n, beta_0, beta_1, beta_2, beta_3){
  X <- sample(c(0, 1), n, replace = TRUE, prob = c(0.3, 0.7))
  W <- sample(c(0, 1), n, replace = TRUE, prob = c(0.6, 0.4))

  u <- rnorm(n)

  Y1 <- beta_0 + beta_1*X + beta_2*W + u
  Y2 <- beta_0 + beta_1*X + beta_2*W + beta_3*X*W + u

  df <- data.frame(Y1, Y2, X, W, u)

  return(df)
}
```

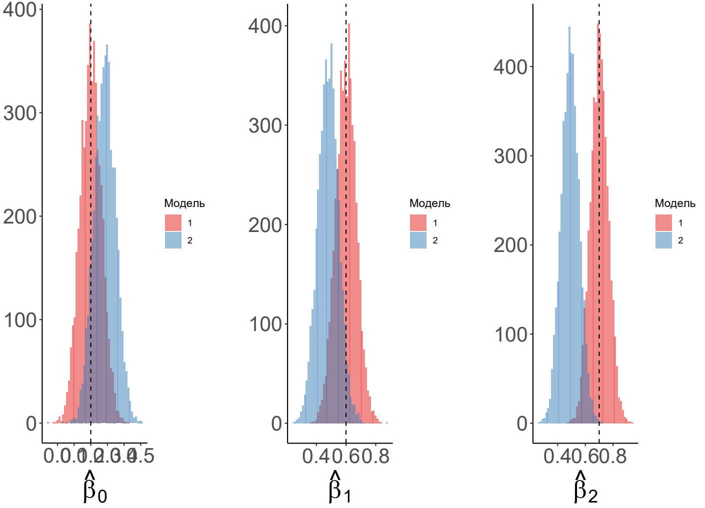
Ілюстрація методом Монте-Карло (5)

- Проводимо саму симуляцію

```r
get_beta <- function(n, beta_0, beta_1, beta_2, beta_3){
  df <- generate_data(n, beta_0, beta_1, beta_2, beta_3)
  model1 <- lm(Y1 ~ X + W, data = df)
  model2 <- lm(Y2 ~ X + W, data = df)

  result <- c(model1$coefficients, model2$coefficients)
  names(result) <- c("Intercept1", "X1", "W1", "Intercept2", "X2", "W2")

  return(result)
}

set.seed(100)
T <- 5000
n <- 1000
beta_0 <- 0.2
beta_1 <- 0.6
beta_2 <- 0.7
beta_3 <- -0.3

df <- as_tibble(t(replicate(T, get_beta(n, beta_0, beta_1, beta_2, beta_3))))
df <- df %>% pivot_longer(cols = everything(),
                          names_pattern = "(.*)(\\d)",
                          names_to = c(".value", "model"))
```

Ілюстрація методом Монте-Карло (6)

- Відповідні гістограми:



- Бачимо наявність зміщення, величина якого повністю відповідає нашим розрахункам

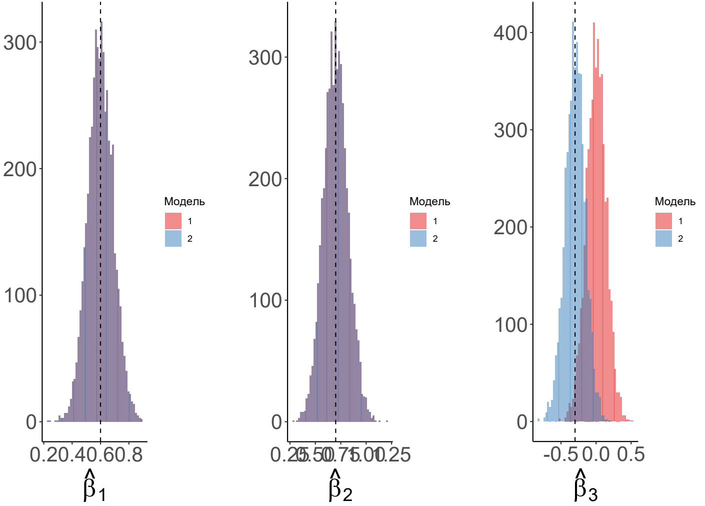
Ілюстрація методом Монте-Карло (7)

- Проводимо тепер симуляцію, де для обох моделей оцінімо проєкцію з фактором взаємодії $X_i W_i$

```
get_beta <- function(n, beta_0, beta_1, beta_2, beta_3){
  df <- generate_data(n, beta_0, beta_1, beta_2, beta_3)
  model1 <- lm(Y1 ~ X + W + I(X*W), data = df)
  model2 <- lm(Y2 ~ X + W + I(X*W), data = df)

  result <- c(model1$coefficients, model2$coefficients)
  names(result) <- c("Intercept1", "X1", "W1", "XW1", "Intercept2", "X2", "W2", "XW2")

  return(result)
}

set.seed(100)
T <- 5000
n <- 1000
beta_0 <- 0.2
beta_1 <- 0.6
beta_2 <- 0.7
beta_3 <- -0.3

df <- as_tibble(t(replicate(T, get_beta(n, beta_0, beta_1, beta_2, beta_3))))
df <- df %>% pivot_longer(cols = everything(),
                          names_pattern = "(.*)(\\d)",
                          names_to = c(".value", "model"))
```

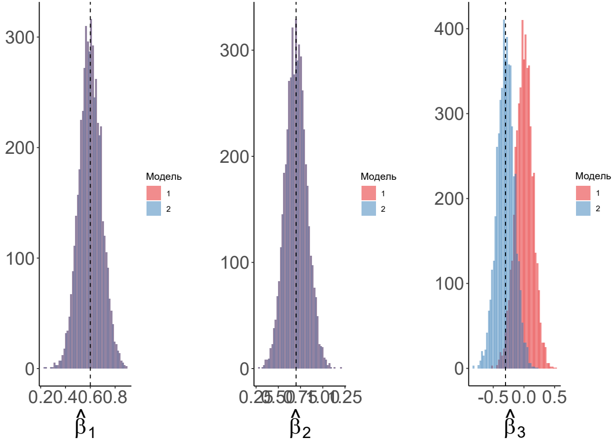
Ілюстрація методом Монте-Карло (8)

- Відповідні гістограми:



- Як можна бачити, в обох випадках оцінки $\widehat{\beta}_1$ і $\widehat{\beta}_2$ є не те що незміщеними, а однаковими

Ілюстрація методом Монте-Карло (9)



- До того ж для моделі 1 коефіцієнт біля $X_i W_i$ прямує до 0
  - Тобто додавання нових змінних, **яких у справжній моделі** не було, не впливає на оцінки інших коефіцієнтів
  - Але додавання таких змінних необмежено не є доречним, адже він цього **зростає дисперсія** оцінок
  - Про це поговоримо трішки згодом

Контрольні змінні (1)

- Отже ми усвідомили всі ризики оцінювання коефіцієнтів **структурної** моделі $Y_i = \mathbf{X}_i^\top \beta + e_i$ за допомогою OLS, якщо e_i корелює з (принаймні деякими змінними з) $\mathbf{X}_i$
    - У цьому випадку з'являється OVB
- Тому однією з можливостей прибрати OVB є включення в модель змінних, які корелюють з іншими (ендогенними) змінними
    - Зрозуміло, що такий підхід більше нагадує «мистецтво», ніж «науку»
    - Існують і інші, які розглядатимемо пізніше в нашому курсі
- Змінні, які ми явно включаємо в модель, щоб зменшити OVB, називають **контрольними** (control variables)

Контрольні змінні (2)

- Отже, маємо модель
$$Y_i = \mathbf{X}_i^\top \beta + \mathbf{W}_i^\top \delta + \tilde{e}_i$$
    - Тут $\mathbf{W}$ — контрольні змінні
    - Ми вважаємо, що $\mathrm{Cov}\left(\tilde{e}_i, \mathbf{X}_i\right) = 0$
- Звісно, можна далі продовжувати ці міркування і казати, що може таке бути, що $\mathrm{Cov}\left(\tilde{e}_i, \mathbf{W}_i\right) \neq 0$
    - А відтак оцінки OLS $\widehat{\delta}$ не будуть спроможні
- Проте всьому є межа
- Контрольні змінні самі по собі інтересу для дослідника **не мають**
    - Нас не цікавить, чи будуть оцінки OLS $\widehat{\delta}$ спроможні
    - Головне, що **тепер** оцінки OLS $\widehat{\beta}$ будуть спроможні

Контрольні змінні (3)

- Трішки формальніше, ми кажемо про те, що замість умови $\mathbb{E}\left[\tilde{e}_i \mid \mathbf{X}_i\right] = 0$ ми вимагаємо виконання $\mathbb{E}\left[\tilde{e}_i \mid \mathbf{X}_i, \mathbf{W}_i\right] = \mathbb{E}\left[\tilde{e}_i \mid \mathbf{W}_i\right]$
- Справді, тоді можемо розглянути $v_i = \tilde{e}_i - \mathbb{E}\left[\tilde{e}_i \mid \mathbf{X}_i, \mathbf{W}_i\right]$
  - Тоді $\mathbb{E}\left[v_i \mid \mathbf{X}_i, \mathbf{W}_i\right] = \mathbb{E}\left[\tilde{e}_i \mid \mathbf{X}_i, \mathbf{W}_i\right] - \mathbb{E}\left[\tilde{e}_i \mid \mathbf{X}_i, \mathbf{W}_i\right] = 0$
- Маємо:

$$\begin{aligned} Y_i &= \mathbf{X}_i^\top \beta + \mathbf{W}_i^\top \delta + \tilde{e}_i = \mathbf{X}_i^\top \beta + \mathbf{W}_i^\top \delta + \mathbb{E}\left[\tilde{e}_i \mid \mathbf{X}_i, \mathbf{W}_i\right] + v_i \\ &= \mathbf{X}_i^\top \beta + \mathbf{W}_i^\top \delta + \mathbb{E}\left[\tilde{e}_i \mid \mathbf{W}_i\right] + v_i \end{aligned}$$

- У цій моделі $\mathbb{E}\left[v_i \mid \mathbf{X}_i, \mathbf{W}_i\right] = 0$, а відтак коефіцієнти біля $\mathbf{X}_i$ буде оцінено спроможною оцінкою як коефіцієнтів лінійної проєкції $\widehat{\beta}$
  - Біля $\mathbf{W}_i$ буде стояти щось зовсім інше, залежно від того, чому дорівнює $\mathbb{E}\left[\tilde{e}_i \mid \mathbf{W}_i\right]$
  - Але воно нас і **не цікавить**
- У цьому випадку коефіцієнти β будуть мати причиново-наслідкову інтерпретацію для $\mathbf{X}_i$

Зв'язок із рандомізованими й контрольованими дослідженнями (1)

- Розгляньмо найпростішу модель $Y_i = \beta_0 + \beta_1 X_i + e_i$
- Вимога $\mathbb{E}\left[e_i \mid X_i\right] = 0$ досягатиметься, зокрема, якщо $e_i \perp\!\!\!\perp X_i$, тобто якщо значення змінної X_i не залежить від жодних інших факторів
- Це є характеристичною особливістю **рандомізованого контрольованого дослідження** (randomized controlled trial, RCT)
  - Наприклад, ми ділимо всіх пацієнтів на дві групи — групу лікування (treatment group) і контрольну (control group)
  - Першій групі дають ліки ($X_i = 1$), а другій — плацебо ($X_i = 0$)
  - **Якщо** ліки і плацебо роздаються абсолютно випадково, то $\mathbb{E}\left[e_i \mid X_i\right] = 0$ за побудовою
- На практиці ми часто стикаємося з даними **спостережень** (observational data), які не є результатом проведення RCT
- Тому вимога $\mathbb{E}\left[e_i \mid X_i\right] = 0$ має таку інтерпретацію, що X_i «нічим не гірше від рандомного» (або, як кажуть, as good as randomly assigned)

Зв'язок із рандомізованими й контрольованими дослідженнями (2)

- Повернімося до нашого прикладу з каліфорнійськими школами
- Ми могли б суто гіпотетично у **повністю** випадковий спосіб розподіляти учнів між класами різного розміру
- Проте на практиці це нереалістично
- Цілком очевидно, що на бали за тести впливають не тільки розмір класу чи навіть кількість неангломовних, а й **зовнішні** фактори
  - Наприклад, доходи родини (це може означати доступ до ліпших засобів навчання)
  - Або благополуччя району (від цього залежить якість школи)
  - Такі параметри корелюють із розміром класу (швидше за все, від'ємно)
- Проте ці зовнішні фактори так чи інакше корелюють із економічним становищем учня
  - Його можна до певної міри поміряти
  - Зокрема, в датасеті є змінна `meal_pct`, яка показує (у відсотках!) частку учнів, які отримують безкоштовні обіди
  - Оскільки на них претендують малозабезпечені сім'ї (приблизно 150% від межі бідності), це є показник економічного становища в шкільному окрузі

Зв'язок із рандомізованими й контрольованими дослідженнями (3)

- Чому це повинно спрацювати?
- Ми можемо розглянути модель

$$Y_i = \beta_0 + \beta_1 X_i + \beta_2 V_i + \beta_3 W_i + e_i$$

  - Тут $X_i =$`str`, $V_i =$`el_pct`, $W_i =$`meal_pct`
  - Це буде структурна модель чи лінійна проєкція?!
- Ми можемо аргументувати, що $\mathbb{E}\left[e_i \mid X_i, V_i, W_i\right] = \mathbb{E}\left[e_i \mid V_i, W_i\right]$
  - Тобто що для округів з однаковою часткою «бідних» студентів та неангломовних студентів розмір класу є «нічим не гіршим від випадкового»
  - Іншими словами, для шкіл з однаковими значеннями V_i і W_i значення X_i можна вважати повністю випадковими
- Тоді можемо говорити про причиново-наслідковий вплив розміру класу X_i, **контролюючи** V_i і W_i
- Але з цієї моделі **не можна казати** про, наприклад, причиново-наслідковий вплив частки бідних студентів W_i
  - Оскільки вона корелює з іншими неврахованими характеристиками типу матеріально-технічного забезпечення школи в окрузі
  - А відтак не можна казати, що W_i «випадкова»
- Звісно, переконливість саме такої аргументації залежить від багатьох факторів, що робить аналіз даних доволі непростою і творчою справою!

Зв'язок із рандомізованими й контрольованими дослідженнями (4)

- Конкретні оцінки для нашого датасету дорівнюють

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

```
##
## Call:
## lm(formula = testscr ~ str + el_pct + meal_pct, data = caschool)
##
## Coefficients:
## (Intercept)          str       el_pct     meal_pct
##    700.1500      -0.9983      -0.1216      -0.5473
```

- Ми бачимо, що значення коефіцієнта біля `str` змінилася несуттєво порівняно з попереднім випадком
- Також ми бачимо, що на перший погляд збільшення частки бідних студентів з 0 до 50% веде до зменшення середнього балу на 27.367 балів
  - Але сприймати це як причиново-наслідковий результат було б просто **дебілізмом**!
  - Адже тоді потрібно просто скасувати всі безкоштовні обіди, і рівень знань автоматично підскоче!
- У той же час коефіцієнт біля `str` можна, хоча б із певною натяжкою, вважати причиново-наслідковим
  - Звісно, додавання інших змінних або маніпуляції з функційною формою рівняння (додати квадрати, логарифми тощо) може ще поліпшити наші результати

Мультиколінеарність (1)

- **Мультиколінеарність** (multicollinearity) постає тоді, коли дві або більше залежних змінних у моделі сильно корельовані між собою
- Розрізняють **повну** (perfect) та **часткову** (imperfect) мультиколінеарності
- Повна мультиколінеарність має місце, коли один із регресорів є лінійною комбінацією інших
- Як правило, це свідчить про некоректність самої модели
- У цьому випадку матриця $\mathbf{X}^\top \mathbf{X}$ не має оберненої, і OLS-оцінка не існує

Мультиколінеарність (2)

- Типовий приклад, коли може виринати мультиколінеарність — наявність категорійного регресора
- Як ми говорили минулої лекції, просто так взяти й додати таку змінну в модель не має сенсу
  - Навіть якщо категорії впорядковані, між ними не визначено поняття відстани
- Тому стандартною практикою є включення в модель *окремих* індикаторних змінних, які відповідають *окремим* категоріям
- Наприклад, нас можуть цікавити продажі морозива в різних локаціях у різні квартали року
  - Тоді кожне спостереження i можна розглядати як окрему локацію
- Щоб з'ясувати вплив кварталу року на продажі, потрібно розглянути змінні $Q_{j,i}$, $j = 1, 2, 3, 4$ — для кожного кварталу
  - Якщо $Q_{j,i} = 1$, то продаж у локації i стався в квартал j
- Тоді модель могла б бути такою:

$$Y_i = \beta_0 + \beta_1 Q_{1,i} + \beta_2 Q_{2,i} + \beta_3 Q_{3,i} + \beta_4 Q_{4,i} + u_i$$

- Проте в такій моделі наявна мультиколінеарність, адже $Q_{1,i} + Q_{2,i} + Q_{3,i} + Q_{4,i} = 1$, що є лінійною функцією від вектора констант
  - Або іншими словами $Q_{4,i} = 1 - (Q_{1,i} + Q_{2,i} + Q_{3,i})$, тобто змінні лінійно залежні
- В англомовній літературі таку ситуацію називають **пасткою індикаторних змінних** (dummy variable trap)

Мультиколінеарність (3)

- Щоб позбутися повної мультиколінеарности, потрібно викинути змінні, які є лінійно залежними
- Якщо R, намагаючись підрахувати OLS-коефіцієнти, зіткнеться з мультиколінеарністю, він викине лінійно залежні змінні на власний розсуд
  - Тому ліпше аналізувати відповідну ситуацію та визначати базову категорію самостійно
- Наприклад, додаймо в нашу модель зі школами, на доданок до частки неангломовних учнів частку англомовних

```
caschool <- caschool %>% mutate(nel_pct = 100 - el_pct)
model <- lm(testscr ~ str + el_pct + nel_pct + meal_pct, data = caschool)
model
```

```
##
## Call:
## lm(formula = testscr ~ str + el_pct + nel_pct + meal_pct, data = caschool)
##
## Coefficients:
## (Intercept)          str       el_pct      nel_pct     meal_pct
##    700.1500      -0.9983      -0.1216           NA      -0.5473
```

Мультиколінеарність (4)

- У контексті пастки індикаторних змінних потрібно відкинути індикаторну змінну, яка відповідає котрійсь категорії
  - Таку категорію називають **базовою**
  - Значення коефіцієнтів для інших категорій тоді будуть мати інтепретацію, відносну до базової категорії
- У нашому прикладі, якщо викинути четвертий квартал, дістанемо таку модель:

$$Y_i = \beta_0 + \beta_1 Q_{1,i} + \beta_2 Q_{2,i} + \beta_3 Q_{3,i} + u_i$$

- Якщо в цій моделі $\mathbb{E}\left[u_i \mid Q_{1,i}, Q_{2,i}, Q_{3,i}\right] = 0$, то β_0 можна інтепретувати як $\mathbb{E}\left[Y_i \mid Q_{4,i} = 1\right] = \mathbb{E}\left[Y_i \mid Q_{1,i} = 0, Q_{2,i} = 0, Q_{3,i} = 0\right]$
- Аналогічно,

$$\mathbb{E}\left[Y_i \mid Q_{1,i} = 1\right] = \mathbb{E}\left[Y_i \mid Q_{1,i} = 1, Q_{2,i} = 0, Q_{3,i} = 0\right] = \beta_0 + \beta_1$$

  - Звідси $\beta_1 = \mathbb{E}\left[Y_i \mid Q_{1,i} = 1\right] - \mathbb{E}\left[Y_i \mid Q_{4,i} = 1\right]$
- Аналогічно $\beta_2 = \mathbb{E}\left[Y_i \mid Q_{2,i} = 1\right] - \mathbb{E}\left[Y_i \mid Q_{4,i} = 1\right]$, $\beta_3 = \mathbb{E}\left[Y_i \mid Q_{3,i} = 1\right] - \mathbb{E}\left[Y_i \mid Q_{4,i} = 1\right]$

Мультиколінеарність (5)

- Якщо маємо неповну мультиколінеарність, деякі регресори сильно корелюють між собою, але не стовідсотково
- Неповна мультиколінеарність є особливістю даних, а не моделі
- Наприклад, якщо вибірка доволі гомогенна, значення в змінних матимуть малий розкид
- У цьому випадку матриця $\mathbf{X}^\top\mathbf{X}$ *має* обернену
- Проте OLS-оцінки будуть не такі точні і сильно залежатимуть від шуму в даних

План лекції

1. Зміщення від неврахованих змінних

2. Статистичне виведення для коефіцієнтів регресії

Спроможність та асимптотичний розподіл OLS-оцінки (1)

- Варто коротко пригадати, як ми виводимо властивості OLS-оцінки
- Для того, щоб їх вивести, потрібно зробити декілька **припущень**
- Маємо модель $Y_i = \mathbf{X}_i^\top \beta + u_i$
- **Припущення 1**: умовне сподівання похибок дорівнює 0: $\mathbb{E}\left[u_i \mid \mathbf{X}_i\right] = 0$
  - Або, якщо $Y_i = \mathbf{X}_i^\top \beta + \mathbf{W}_i^\top \delta + u_i$, що $\mathbb{E}\left[u_i \mid \mathbf{X}_i, \mathbf{W}_i\right] = \mathbb{E}\left[u_i \mid \mathbf{W}_i\right]$
- **Припущення 2**: $(X_{1,i}, \ldots, X_{k,i}, Y_i)$, $i = 1, \ldots, n$ незалежні та мають однаковий розподіл
- **Припущення 3**: $0 < \mathbb{E}\left[X_{j,i}^4\right] < \infty\ ,\ j = 1, \ldots, k\ , \quad 0 < \mathbb{E}\left[Y_i^4\right] < \infty$
- **Припущення 4**: матриця $\mathbf{X}^\top \mathbf{X}$ має обернену

Спроможність та асимптотичний розподіл OLS-оцінки (2)

- Розпишімо Y_i:

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

- Згідно з ЗВЧ,

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

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

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

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

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

Спроможність та асимптотичний розподіл OLS-оцінки (3)

- 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 $\mathbf{X}_i u_i$, $i=1,\ldots,n$, незалежні, можемо застосувати ЦГТ до другого множника:

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

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

$$\sqrt{n}\left(\hat{\beta}-\beta\right)\xrightarrow{d}N\left(0,\left(\mathbb{E}\left[\mathbf{X}_i\mathbf{X}_i^{\top}\right]\right)^{-1}\mathbb{E}\left[\mathbf{X}_i\mathbf{X}_i^{\top}u_i^2\right]\left(\mathbb{E}\left[\mathbf{X}_i\mathbf{X}_i^{\top}\right]\right)^{-1}\right)\equiv N\left(0,\mathbf{V}_{\beta}\right) \tag{2.1}$$

Обчислення стандартних похибок OLS (1)

- Для побудови довірчих інтервалів та тестування гіпотез із коефіцієнтами OLS потрібно знати їхні стандартні похибки
- Для цього потрібно знайти оцінку асимптотичної дисперсії
- Plug-in оцінкою **могла б бути**

$$\hat{\mathbf{V}}_\beta = \left(\frac{1}{n} \sum_{i=1}^{n} \mathbf{X}_i \mathbf{X}_i^\top \right)^{-1} \left(\frac{1}{n} \sum_{i=1}^{n} \mathbf{X}_i \mathbf{X}_i^\top u_i^2 \right) \left(\frac{1}{n} \sum_{i=1}^{n} \mathbf{X}_i \mathbf{X}_i^\top \right)^{-1}$$

- Проте цю «оцінку» неможливо використати, бо вона залежить від **невідомих** u_i
- Відтак на практиці невідомі похибки u_i замінюють на відповідні **залишки** $\hat{u}_i = Y_i - \hat{Y}_i = Y_i - \mathbf{X}_i^\top \hat{\beta}$
- Тоді дістаємо таку оцінку:

$$\hat{\mathbf{V}}_\beta^{HC0} = \left(\frac{1}{n} \sum_{i=1}^{n} \mathbf{X}_i \mathbf{X}_i^\top \right)^{-1} \left(\frac{1}{n} \sum_{i=1}^{n} \mathbf{X}_i \mathbf{X}_i^\top \hat{u}_i^2 \right) \left(\frac{1}{n} \sum_{i=1}^{n} \mathbf{X}_i \mathbf{X}_i^\top \right)^{-1} \qquad (2.2)$$

  - Її позначають **HC0** (від англійського *heteroskedasticity-consistent*)
  - Можна показати, що вона є спроможною, проте це виходить за рамки нашого курсу
  - Подібні оцінки часто називають **сендвічними** (sandwich)

- Минулої лекції ми згадували, що

$$\hat{\sigma}_u^2 = \frac{1}{n-k-1}\sum_{i=1}^{n}\hat{u}_i^2$$

є незміщеною оцінкою дисперсії похибок Var (u_i)

- За цією ж логікою на практиці **рекомендується** використовувати таку оцінку дисперсії:

$$\begin{aligned}\hat{\mathbf{V}}_\beta^{HC1} &= \left(\frac{1}{n}\sum_{i=1}^{n}\mathbf{X}_i\mathbf{X}_i^\top\right)^{-1}\left(\frac{1}{n-k-1}\sum_{i=1}^{n}\mathbf{X}_i\mathbf{X}_i^\top\hat{u}_i^2\right)\left(\frac{1}{n}\sum_{i=1}^{n}\mathbf{X}_i\mathbf{X}_i^\top\right)^{-1} \\ &= \frac{n}{n-k-1}\hat{\mathbf{V}}_\beta^{HC0}\end{aligned} \tag{2.3}$$

  - Існують іще версії HC2, HC3 та HC4, але вони використовуються значно рідше

- У будь-якому випадку, зверніть увагу, що $\sqrt{n}\left(\hat{\beta}-\beta\right)\xrightarrow{d}N\left(0,\mathbf{V}_\beta\right)$, тобто $\hat{\beta}\overset{\text{a}}{\sim}N\left(\beta,\frac{1}{n}\mathbf{V}_\beta\right)$!
  - Відтак стандартні похибки окремих коефіцієнтів дорівнюють $\widehat{\text{se}}\left(\hat{\beta}_j\right)=\sqrt{\frac{\hat{\mathbf{V}}_{\beta,jj}}{n}}$!

Обчислення стандартних похибок OLS в R (1)

- Як зазначалося на минулій лекції, функція `summary`, застосована до результату застосування функції `lm`, дає в тому числі інформацію про стандартні похибки коефіцієнтів OLS

```r
model <- lm(testscr ~ str + el_pct + meal_pct, data = caschool)
summary(model)
```

```
##
## Call:
## lm(formula = testscr ~ str + el_pct + meal_pct, data = caschool)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -32.849  -5.151  -0.308   5.243  31.501
##
## Coefficients:
##              Estimate Std. Error t value Pr(>|t|)
## (Intercept) 700.14997    4.68569 149.423  < 2e-16 ***
## str          -0.99831    0.23875  -4.181 3.54e-05 ***
## el_pct       -0.12157    0.03232  -3.762 0.000193 ***
## meal_pct     -0.54735    0.02160 -25.341  < 2e-16 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 9.08 on 416 degrees of freedom
## Multiple R-squared:  0.7745, Adjusted R-squared:  0.7729
## F-statistic: 476.3 on 3 and 416 DF,  p-value: < 2.2e-16
```

- Стандартні похибки містяться в колонці `Std. Error`

Обчислення стандартних похибок OLS в R (2)

- Проблема полягає в тому, що ці стандартні похибки **за замовчуванням** є гомоскедастичними
  - Це нам зовсім непотрібно
- Для застосування оцінок HC0 чи HC1 потрібно використати функцію `coeftest` з пакета `lmtest`
- Потрібно додатково вказати, яку оцінку дисперсії ми використовуємо, за допомогою функції `hccm` з пакета `car`

```
coeftest(model, vcov. = hccm(model, type = "hc1"))
```

```
##
## t test of coefficients:
##
##                Estimate Std. Error  t value  Pr(>|t|)
## (Intercept) 700.149966   5.568450 125.7352 < 2.2e-16 ***
## str          -0.998309   0.270080  -3.6963 0.0002480 ***
## el_pct       -0.121573   0.032832  -3.7029 0.0002418 ***
## meal_pct     -0.547346   0.024107 -22.7046 < 2.2e-16 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

- Для того, щоб можна було наочно побачити відмінності між різними способами оцінки стандартних похибок, можна використати дуже корисну функцію `stargazer` з однойменного пакета

```
model_hc0 <- coeftest(model, vcov. = hccm(model, type = "hc0"))
model_hc1 <- coeftest(model, vcov. = hccm(model, type = "hc1"))

stargazer(model, model_hc0, model_hc1,
          type = "text",
          digits = 3)
```

```
##
## ==================================================================
##                                 Dependent variable:
##                       --------------------------------------------
##                               testscr
##                                 OLS               coefficient
##                                                      test
##                                 (1)              (2)        (3)
## ------------------------------------------------------------------
## str                         -0.998***        -0.998***  -0.998***
##                              (0.239)          (0.269)    (0.270)
##
## el_pct                      -0.122***        -0.122***  -0.122***
##                              (0.032)          (0.033)    (0.033)
##
## meal_pct                    -0.547***        -0.547***  -0.547***
##                              (0.022)          (0.024)    (0.024)
##
## Constant                    700.150***       700.150*** 700.150***
##                              (4.686)          (5.542)    (5.568)
##
## ------------------------------------------------------------------
## Observations                  420
## R2                           0.775
## Adjusted R2                  0.773
## Residual Std. Error     9.080 (df = 416)
## F Statistic         476.306*** (df = 3; 416)
```

Обчислення стандартних похибок OLS в R (4)

- Ми вказали базову модель та дві варіації з різними гетероскедастичними похибками
- Також ми вказали, що результат хочемо бачити в текстовому форматі
  - Інші опції — формат LaTeX (`type = "latex"`) та HTML (`type = "html"`)
- Можемо бачити, що гетероскедастичні похибки більші
  - Як правило, так і буває
  - Гомоскедастичні похибки в цьому сенсі особливо «небезпечні», бо можуть створити хибне враження, ніби той чи той коефіцієнт є статистично значущий

Побудова довірчих інтервалів для OLS-коефіцієнтів

- Оскільки OLS-коефіцієнти $\widehat{\beta}$ мають асимптотичний нормальний розподіл, для них можна збудувати довірчі інтервали з покриттям $1-\alpha$ за стандартною формулою:

$$C_{\hat{\beta}} = \left[\widehat{\beta}_j - z_{\alpha/2}\widehat{\mathrm{se}}\left(\widehat{\beta}_j\right); \widehat{\beta}_j + z_{\alpha/2}\widehat{\mathrm{se}}\left(\widehat{\beta}_j\right)\right]$$

- В R це можна зробити за допомогою функції `coefci` з пакета `lmtest`
- Зокрема, для 95% інтервалів маємо:

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

```
##                   2.5 %       97.5 %
## (Intercept) 689.2041596 711.09577284
## str          -1.5292007  -0.46741791
## el_pct       -0.1861100  -0.05703661
## meal_pct     -0.5947328  -0.49995834
```

- На практиці в дослідженнях публікують значення коефіцієнтів та відповідних стандартних похибок
    - Тоді читач може швидко зорієнтуватися щодо 95% інтервалів, додавши до $\widehat{\beta}_j$ «плюс-мінус» $2\widehat{\mathrm{se}}\left(\widehat{\beta}_j\right)$

Тестування гіпотез щодо одного коефіцієнта (1)

- Оскільки кожний $\hat{\beta}_j \overset{\text{a}}{\sim} N\left(\beta_j, \left(\text{se}\left(\hat{\beta}_j\right)\right)^2\right)$, для тестування гіпотез щодо значень окремих коефіцієнтів потрібно використовувати **тест Волда**
  - На практиці все одно використовують t-розподіли замість стандартних нормальних, хоча на великих вибірках різниці немає
  - Понад те, якщо похибки гетероскедастичні, то на малих вибірках використання t-розподілу є необґрунтованим[3]!
- Найпоширеніша гіпотеза, яку тестують на практиці — $H_0 : \beta_j = 0$ vs. $H_1 : \beta_j \neq 0$
  - Тобто нас цікавить, має X_j вплив на Y чи не має
  - Якщо модель має причиново-наслідкову інтерпретацію, то цей результат дуже важливий для вироблення практичних рішень
  - Якщо ж ні, то принаймні можна оцінити наявність чи відсутність *статистичного зв'язку* між змінними

[3] *Econometrics* (Hansen), p. 146

Тестування гіпотез щодо одного коефіцієнта (2)

- Функція `coeftest` в R здійснює цей тест автоматично, для кожного коефіцієнта:

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

```
##
## t test of coefficients:
##
##                Estimate Std. Error  t value  Pr(>|t|)
## (Intercept) 700.149966   5.568450 125.7352 < 2.2e-16 ***
## str          -0.998309   0.270080  -3.6963 0.0002480 ***
## el_pct       -0.121573   0.032832  -3.7029 0.0002418 ***
## meal_pct     -0.547346   0.024107 -22.7046 < 2.2e-16 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

- Колонка `t value` містить значення відповідних тестових статистик, які для кожного коефіцієнта β_j дорівнюють

$$T_j = \frac{\widehat{\beta}_j - 0}{\widehat{\text{se}}\left(\widehat{\beta}_j\right)} = \frac{\widehat{\beta}_j}{\widehat{\text{se}}\left(\widehat{\beta}_j\right)}$$

Тестування гіпотез щодо одного коефіцієнта (3)

- Колонка `Pr(>|t|)` містить відповідне p-значення
  - Для його обчислення використовують t-розподіл із $n - k - 1$ ступенями вільности
  - Хоча для великих вибірок було б достатньо й стандартного нормального
- Зірочки привертають увагу дослідника до коефіцієнтів, які є найбільш статистично значущими
  - Хоча, як зазначалося раніше, сучасною практикою є публікація результатів досліджень без цих зірочок
  - Вони можуть приводити до недоречної інтерпретації, що «коефіцієнти з зірочками» є найбільш важливими
  - Насправді важливо дивитися не тільки на статистичну значущість коефіцієнта, а й на його **величину**

Тестування на рівність двох вибіркових середніх (1)

- У попередніх лекціях ми розглядали приклад застосування тесту Волда до тестування різниці вибіркових середніх
- Ми казали, що в нас є дві вибірки $X_1, \dots, X_n \overset{\text{i.i.d.}}{\sim} X$ і $Y_1, \dots, Y_m \overset{\text{i.i.d.}}{\sim} Y$ незалежних між собою величин
- І ми хотіли перевірити гіпотезу $H_0 : \mathbb{E}[X] = \mathbb{E}[Y]$ vs. $H_1 : \mathbb{E}[X] \neq \mathbb{E}[Y]$
- Виявляється, цю ж гіпотезу можна перевірити й у рамках регресійного аналізу
- Справді, нехай маємо модель $Y_i = \beta_0 + \beta_1 X_i + e_i$, $\mathbb{E}[e_i \mid X_i] = 0$
- Тоді, очевидно, $\mathbb{E}[Y_i \mid X_i = 0] = \beta_0$, $\mathbb{E}[Y_i \mid X_i = 1] = \beta_0 + \beta_1$
- Відтак перевірити рівність $\mathbb{E}[Y_i \mid X_i = 0] = \mathbb{E}[Y_i \mid X_i = 1]$ можна, протестувавши гіпотезу, що $\beta_1 = 0$
  - До того, якщо ця модель справді істинна, то це матиме причиново-наслідкову інтерпретацію
  - Це значно потужніше, ніж просто тестувати два вибіркові середні

Тестування на рівність вибіркових середніх (2)

- Проте, навіть якщо ми маємо **всього лише** проєкцію $Y_i = \beta_0 + \beta_1 X_i + u_i$, $\mathbb{E}\left[X_i u_i\right] = 0$, ми все одно можемо протестувати гіпотезу про рівність вибіркових середніх
- Справді, позначмо $\overline{Y}_0 = \dfrac{\sum_{i=1}^n Y_i \mathbb{1}\left\{X_i = 0\right\}}{\sum_{i=1}^n \mathbb{1}\left\{X_i = 0\right\}}$
    - Тобто це вибіркове середнє підгрупи, для якої $X_i = 0$
    - Аналогічно позначмо $\overline{Y}_1$
- Тоді оцінка OLS розв'язує задачу

$$\begin{aligned}\left(\hat{\beta}_0, \hat{\beta}_1\right) &= \operatorname*{arg\,min}_{b_0, b_1} \sum_{i=1}^n \left(Y_i - b_0 - b_1 X_i\right)^2 \\ &= \operatorname*{arg\,min}_{b_0, b_1} \sum_{i: X_i = 0} \left(Y_i - b_0\right)^2 + \sum_{i: X_i = 1} \left(Y_i - b_0 - b_1\right)^2\end{aligned}$$

Тестування на рівність вибіркових середніх (3)

- Уведімо позначення $\gamma = \beta_0 + \beta_1$:

$$\left(\widehat{\beta}_0, \widehat{\gamma}\right) = \left(\arg\min_{b_0} \sum_{i:X_i=0} \left(Y_i - b_0\right)^2, \arg\min_{c} \sum_{i:X_i=1} \left(Y_i - c\right)^2\right)$$

- Тоді умова першого порядку для першої задачі дає $-2\sum_{i:X_i=0}(Y_i - b_0) = 0$
  - Звідси $\widehat{\beta}_0 = \overline{Y}_0$
- Аналогічно $\widehat{\gamma} = \widehat{\beta}_0 + \widehat{\beta}_1 = \overline{Y}_1$
  - Звідси $\widehat{\beta}_1 = \overline{Y}_1 - \overline{Y}_0$
- Тобто регресійну модель, навіть якщо вона не має причиново-наслідкової інтерпретації, можна використовувати для тестування рівности вибіркових середніх

Тестування на рівність вибіркових середніх (4)

- Понад те, у такий самий спосіб можна розглянути декілька різних груп
- Повернімося до датасету про зарплати працівників США з минулої лекції

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

```
## Rows: 50742 Columns: 12
## -- Column specification --------------------------------------------------------
## Delimiter: ";"
## dbl (12): age, female, hisp, education, earnings, hours, week, union, uncov,...
##
## 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.
```

- Розгляньмо регіони, яких у датасеті є чотири
- Нехай нас цікавить перевірити, чи є різниця в середніх (логаритмах) зарплат між різними регіонами
- Ми можемо робити попарні тести Волда / t-тести між різними регіонами

Тестування на рівність вибіркових середніх (5)

- Це ж можна зробити і за допомогою регресії, увівши змінні:
  - $V1 = 1$, якщо спостереження належить регіону 1
  - $V2 = 1$, якщо спостереження належить регіону 2
  - $V3 = 1$, якщо спостереження належить регіону 3
  - Щоб уникнути мультиколінеарности, ми вважаємо регіон 4 базовою категорією
- Тоді маємо

```
wages <- wages %>% mutate(V1 = (region == 1), V2 = (region == 2), V3 = (region == 3))
model_wages <- lm(log_hourly_wage ~ V1 + V2 + V3, data = wages)
model_wages
```

```
##
## Call:
## lm(formula = log_hourly_wage ~ V1 + V2 + V3, data = wages)
##
## Coefficients:
## (Intercept)       V1TRUE       V2TRUE       V3TRUE
##     2.98075      0.07728     -0.06791     -0.07474
```

- Згадаймо інтерпретацію: $\beta_1 = \overline{Y}_1 - \overline{Y}_4, \beta_2 = \overline{Y}_2 - \overline{Y}_4, \beta_3 = \overline{Y}_3 - \overline{Y}_4$

```
mean(wages$log_hourly_wage[wages$V1 == 0 & wages$V2 == 0 & wages$V3 == 0])
```

```
## [1] 2.980747
```

```
mean(wages$log_hourly_wage[wages$V1 == 1 & wages$V2 == 0 & wages$V3 == 0]) -
  mean(wages$log_hourly_wage[wages$V1 == 0 & wages$V2 == 0 & wages$V3 == 0])
```

```
## [1] 0.077282
```

```
mean(wages$log_hourly_wage[wages$V1 == 0 & wages$V2 == 1 & wages$V3 == 0]) -
  mean(wages$log_hourly_wage[wages$V1 == 0 & wages$V2 == 0 & wages$V3 == 0])
```

```
## [1] -0.06791119
```

```
mean(wages$log_hourly_wage[wages$V1 == 0 & wages$V2 == 0 & wages$V3 == 1]) -
  mean(wages$log_hourly_wage[wages$V1 == 0 & wages$V2 == 0 & wages$V3 == 0])
```

```
## [1] -0.07474234
```

Тестування на рівність вибіркових середніх (6)

- Можемо порахувати стандартні похибки

```
coeftest(model_wages, vcov. = hccm(model_wages, type = "hc1"))
```

```
##
## t test of coefficients:
##
##               Estimate Std. Error  t value  Pr(>|t|)
## (Intercept)  2.9807466  0.0057294 520.2543 < 2.2e-16 ***
## V1TRUE       0.0772820  0.0087675   8.8146 < 2.2e-16 ***
## V2TRUE      -0.0679112  0.0079181  -8.5767 < 2.2e-16 ***
## V3TRUE      -0.0747423  0.0076890  -9.7207 < 2.2e-16 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

- Можемо бачити, що всі коефіцієнти є статистично значущі
  - Тобто різниці в середніх зарплатах між кожним регіоном і регіоном 4 статистично значущі
- А що, якщо ми хочемо протестувати різницю між, скажімо, регіонами 1 і 2?
  - Тоді треба тестувати рівність $\beta_{V1} = \beta_{V2}$
- А якщо хочемо протестувати рівність середніх по **всіх** регіонах одночасно, то треба тестувати $\beta_{V1} = \beta_{V2} = \beta_{V3} = 0$

- Гіпотези типу $\beta_{V1} = \beta_{V2}$ належать до класу гіпотез виду $H_0 : \mathbf{a}^\top \beta = c$ vs. $H_1 : \mathbf{a}^\top \beta \neq c$
    - Тут $\mathbf{a} \in \mathbb{R}^{k+1}$, $c \in \mathbb{R}$ — деякі константи
- Оцінкою цієї величини, зрозуміло, є $\mathbf{a}^\top \hat{\beta}$
    - Вона очевидно є спроможною з застосуванням ТНВ
- Якщо H_0 справді істинна, то $\mathbf{a}^\top \hat{\beta} - c = \mathbf{a}^\top \hat{\beta} - \mathbf{a}^\top \beta$
- Асимптотичний розподіл цієї оцінки можна дістати за теоремою Слуцького

$$\sqrt{n} \left(\mathbf{a}^\top \hat{\beta} - \mathbf{a}^\top \beta \right) = \mathbf{a}^\top \cdot \sqrt{n} \left(\hat{\beta} - \beta \right) \xrightarrow{d} N \left(0, \mathbf{a}^\top \mathbf{V}_\beta \mathbf{a} \right)$$

- Звідси маємо стандартну похибку:

$$\widehat{\text{se}} \left(\mathbf{a}^\top \hat{\beta} \right) = \sqrt{\frac{\mathbf{a}^\top \hat{\mathbf{V}}_\beta \mathbf{a}}{n}}$$

- Далі можна застосовувати тест Волда із тестовою статистикою $\frac{\mathbf{a}^\top \hat{\beta} - c}{\widehat{\text{se}} \left(\mathbf{a}^\top \hat{\beta} \right)}$

Тестування гіпотез про лінійні комбінації коефіцієнтів (2)

- Наприклад, для перевірки гіпотези $\beta_1 = \beta_2$ маємо

$$\beta_{V1} - \beta_{V2} = (0, 1, -1, 0) \cdot \begin{pmatrix} \beta_0 \\ \beta_{V1} \\ \beta_{V2} \\ \beta_{V3} \end{pmatrix}$$

  - Тобто $\mathbf{a}^\top = (0, 1, -1, 0)$, $c = 0$
- Далі справа техніки

Тестування гіпотез про декілька коефіцієнтів (1)

- Нехай тепер ми хочемо протестувати $\beta_{V1} = \beta_{V2} = \beta_{V3} = 0$
  - У такий спосіб ми хочемо показати, що **вся змінна** `region` не має впливу на зарплати
- Фактично ми тестуємо **спільну гіпотезу** $H_0 : \beta_{V1} = 0$ і $\beta_{V2} = 0$ і $\beta_{V3} = 0$ vs. H_1 : *принаймні один* коефіцієнт $\neq 0$
- **Неправильний** підхід полягає в тестуванні трьох гіпотез окремо і відкиданні спільної гіпотези, якщо хоча б один тест не пройде
  - Ми говорили в Лекції 5, що в цьому випадку рівень тесту буде більший від $\alpha = 0.05$
- Тому **правильний** підхід полягає в тестуванні гіпотези такого виду: $H_0 : \mathbf{A}\beta = \mathbf{c}$ vs. $H_0 : \mathbf{A}\beta \neq \mathbf{c}$
  - Тут $\mathbf{A}$ — деяка матриця $q \times (k+1)$ з рангом q
  - $\mathbf{c}$ — деякий q-вимірний вектор констант
- Спроможною оцінкою $\mathbf{A}\beta$ буде $\mathbf{A}\widehat{\beta}$
- Її асимптотичний розподіл буде

$$\sqrt{n}\left(\mathbf{A}\widehat{\beta} - \mathbf{A}\beta\right) \xrightarrow{d} N\left(\mathbf{0}, \mathbf{A}\mathbf{V}_{\beta}\mathbf{A}^{\top}\right)$$

Тестування гіпотез про декілька коефіцієнтів (2)

- Але оскільки це є випадковий вектор розмірности q, ми можемо перейти до скаляру, обчисливши F**-статистику**:

$$F = \frac{1}{q}\left(\hat{\beta}^\top \mathbf{A}^\top - \mathbf{c}^\top\right)\left(\mathbf{A}\frac{1}{n}\hat{\mathbf{V}}_\beta \mathbf{A}^\top\right)^{-1}\left(\mathbf{A}\hat{\beta} - \mathbf{c}\right) \sim F_{q,n-k-1} \tag{2.4}$$

  - Якщо H_0 істинна, F має F-розподіл із q та $n-k-1$ ступенями вільности
  - Ми не будемо виводити, чому саме так — усе це є у відповідній літературі

Тестування гіпотез про декілька коефіцієнтів (3)

- Так само, як ми замість t-розподілу на великих вибірках використовуємо стандартний нормальний, замість $F_{q,n-k-1}$-розподілу ми використовуємо розподіл $F_{q,\infty} = \frac{\chi^2_q}{q}$
- Цей факт ми можемо дозволити собі вивести
- Справді, якщо H_0 істинна, то $\mathbf{A}\hat{\beta} - \mathbf{c} = \mathbf{A}\hat{\beta} - \mathbf{A}\beta = \mathbf{A}\left(\hat{\beta} - \beta\right)$
- І тоді, із застосуванням ТНВ та теореми Слуцького, маємо

$$
\begin{aligned}
F &= \left(\hat{\beta}^\top \mathbf{A}^\top - \mathbf{c}^\top\right)\left(\mathbf{A}\frac{1}{n}\hat{\mathbf{V}}_\beta \mathbf{A}^\top\right)^{-1}\left(\mathbf{A}\hat{\beta} - \mathbf{c}\right) = \\
&= \sqrt{n}\left(\mathbf{A}\left(\hat{\beta} - \beta\right)\right)^\top \left(\mathbf{A}\hat{\mathbf{V}}_\beta \mathbf{A}^\top\right)^{-1} \sqrt{n}\mathbf{A}\left(\hat{\beta} - \beta\right) \\
&\xrightarrow{d} \left(N\left(\mathbf{0}, \mathbf{A}\mathbf{V}_\beta \mathbf{A}^\top\right)\right)^\top \cdot \left(\mathbf{A}\mathbf{V}_\beta \mathbf{A}^\top\right)^{-1} \cdot N\left(\mathbf{0}, \mathbf{A}\mathbf{V}_\beta \mathbf{A}^\top\right) \\
&= \left(N\left(\mathbf{0}, \mathbf{I}\right)\right)^\top \cdot N\left(\mathbf{0}, \mathbf{I}\right) \\
&= \chi^2_q
\end{aligned}
$$

  - Останнє випливає за визначенням розподілу χ^2_q як суми квадратів q стандартних нормальних змінних
  - Коли $q = 1$, маємо $\chi^2_1 = t^2$, тобто попередній випадок є частковим

Тестування гіпотез про декілька коефіцієнтів (4)

- Для прикладу $\beta_{V1} = \beta_{V2} = \beta_{V3} = 0$ маємо

$$\mathbf{A} \begin{pmatrix} \beta_0 \\ \beta_{V1} \\ \beta_{V2} \\ \beta_{V3} \end{pmatrix} = \mathbf{0} \quad \Leftrightarrow \quad \begin{pmatrix} 0 & 1 & 0 & 0 \\ 0 & 0 & 1 & 0 \\ 0 & 0 & 0 & 1 \end{pmatrix} \cdot \begin{pmatrix} \beta_0 \\ \beta_{V1} \\ \beta_{V2} \\ \beta_{V3} \end{pmatrix} = \mathbf{0}$$

- Далі справа техніки

Тестування гіпотез про декілька коефіцієнтів (5)

- В R можна протестувати ці гіпотези можна за допомогою функції `linearHypothesis` із пакета `car`
- Зокрема, гіпотезу $\beta_{V1} = \beta_{V2}$ можна протестувати так:

```
linearHypothesis(model_wages, c("V1TRUE = V2TRUE"),
                 vcov = hccm(model_wages, type = "hc1"))
```

```
## Linear hypothesis test
##
## Hypothesis:
## V1TRUE - V2TRUE = 0
##
## Model 1: restricted model
## Model 2: log_hourly_wage ~ V1 + V2 + V3
##
## Note: Coefficient covariance matrix supplied.
##
##   Res.Df Df      F    Pr(>F)
## 1  50625
## 2  50624  1 285.21 < 2.2e-16 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

- Як можна бачити, ми відкидаємо цю гіпотезу, бо p-значення майже нульове
    - Тобто різниця у вибіркових середніх між регіонами 1 і 2 статистично значуща

Тестування гіпотез про декілька коефіцієнтів (6)

- Гіпотезу $\beta_{V1} = \beta_{V2} = \beta_{V3} = 0$ можна протестувати так:

```
linearHypothesis(model_wages, c("V1TRUE = 0", "V2TRUE = 0", "V3TRUE = 0"),
                 vcov = hccm(model_wages, type = "hc1"))
```

```
## Linear hypothesis test
##
## Hypothesis:
## V1TRUE = 0
## V2TRUE = 0
## V3TRUE = 0
##
## Model 1: restricted model
## Model 2: log_hourly_wage ~ V1 + V2 + V3
##
## Note: Coefficient covariance matrix supplied.
##
##   Res.Df Df      F    Pr(>F)
## 1  50627
## 2  50624  3 137.28 < 2.2e-16 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

- Як можна бачити, ми також відкидаємо цю гіпотезу, бо p-значення майже нульове
  - Тобто вплив змінної `region` статистично відмінний від 0