Федеральное государственное автономное образовательное учреждение высшего образования

НАЦИОНАЛЬНЫЙ ИССЛЕДОВАТЕЛЬСКИЙ УНИВЕРСИТЕТ «ВЫСШАЯ ШКОЛА ЭКОНОМИКИ»

Факультет экономических наук

по направлению подготовки Экономика
образовательная программа «Экономика»

Домашнее задание №2

В рамках курса «Микроэконометрика качественных данных»

Выполнил:
Максим Пешков
Группа:
БЭК181

Москва, 2021

Содержание

О данных

В данной работе мы будем изучать, как различные факторы влияют на зарплату женщин с учетом неслучайного отбора в число занятых.

Обозначения переменных:

- lfp (в работе переименовано в employment_dummy) - бинарная переменная, принимающая значение 1 если женщина работает и 0 - в противном случае (независимая переменная)
- wage - почасовая зарплата женщины (независимая переменная)
- hours - количество проработанных часов за год
- kids5 - количество детей младше 6 лет
- kids618 - количество несовершеннолетних детей старше пяти лет
- age - возраст женщины
- educ (в работе переименовано в education)- число лет, потраченных женщиной на получение образования
- hushrs (в работе переименовано husband_work_hours) - количество часов, проработанных мужем женщины
- husage - возраст мужа женщины
- huseduc - число лет, потраченных мужем женщины на получение образования
- huswage - зарплата мужа женщины
- faminc - доход семьи женщины
- mtr - налоговая нагрузка на женщину
- fatheduc - число лет, потраченных отцом женщины на получение образования
- motheduc - число лет, потраченных матерью женщины на получение образования
- unem (в работе переименовано в unemployment_region) - безработица в регионе проживания женщины
- city - бинарная переменная, принимающая значение 1 если женщина живет в городе и 0 - иначе
- exper (в работе переименовано experience) - рабочий стаж женщины в годах
- nwifeinc - доход семьи женщины за вычетом ее дохода
- wifecoll - бинарная переменная, принимающая значение 1 если женщина посещала колледж и 0 - иначе
- huscoll - бинарная переменная, принимающая значение 1 если муж женщины посещал колледж и 0 - иначе

Часть 1. Теория и гипотезы

Задание №1.1.

Выберите независимые переменные для уравнения зарплаты и уравнения занятости. Кратко теоретически обоснуйте выбор каждой из них: не обязательно со ссылками на литературу, достаточно здравого смысла. Укажите и кратко обоснуйте предполагаемые направления эффектов. Уравнение занятости должно включать по крайней мере одну переменную, которой не было в уравнении зарплаты, и одну переменную, которая есть в уравнении зарплаты. Желательно, чтобы общая для двух уравнений переменная была непрерывной, например, возраст, а также не использовать более трех переменных в каждом из уравнений.

Для уравнения зарплаты и зависимой переменной **wage** будем использовать следующие независимые переменные:

- husband_work_hours

  Количество часов, проработанных мужем, отражает несколько вещей: работает ли муж, как усердно работает, сколько может примерно зарабатывать (если считать, что все примерно одинаково зарабатывают, то можно примерно оценить среднюю возможную зарплату). Таким образом, учитывая, что для проживания семьям достаточно получать определенный средний доход, эти факторы позволяют оценить должна ли работать жена (если муж недостаточно работает или не работает вовсе), и если нужно работать, то с какими усилиями нужно, чтобы обеспечить семью достаточным для счастливого проживания общим доходом. То есть если муж мало работает, то им может не хватать денег, поэтому жена будет занимать высокооплачиваемую должность. И наоборот, если муж усердно работает, то скорей всего, у их семьи достаточный общий доход и у жены нет стимулов к усердной работе с высокой зарплатой, что также подкрепляется домашними заботами в семье, которые перекладываются на жену, ведь муж много работает.

  Следовательно, будем предполагать, что количество часов, проработанных мужем, негативно влияет на зарплату жены.

- education

  Существует множество исследований, показывающих, что более высокое образование положительно влияет на уровень зарплаты, так как работник обладает большим количеством знаний и навыков (более высокий человеческий капитал), что ценят на более оплачиваемых должностях и в более крупных компаниях. При этом для женщин этот эффект скорей всего более выраженный, так как в силу естественных обстоятельств (рождение детей) многие теряют 2-3 года на получение более высокого уровня образования, поэтому разница в уровне образования у девушек более выраженная, аналогично и в уровне зарплаты наблюдается более выраженный разрыв.

  Следовательно, будем предполагать, что количество лет, потраченных женщиной на получение образования положительно влияет на зарплату жены.

- experience

  Аналогично уровню образования рабочий стаж женщины в годах может отображать как много знаний и умений работница получила во время работы, ведь в рабочее время все работники улучшают свои профессиональные навыки, получают больше знаний (увеличивают свой человеческий капитал), что способствует их более уверенному продвижению по карьерной лестнице, увеличивая себе уровень зарплаты.

  Следовательно, будем предполагать, что рабочий стаж женщины в годах положительно влияет на зарплату жены.

Для уравнения занятости и зависимой переменной **employment_dummy** будем использовать следующие независимые переменные:

- husband_work_hours

  Аналогично уровню зарплаты количество часов, проработанных мужем, может демонстрировать необходимость жены работать вовсе, ведь если муж много работает, то скорей всего, доход их семьи достаточно высокий и у жены нет высокий стимулов к работе, к тому же, домашние заботы и уход за детьми в таких условиях только уменьшают желание и возможность выходить на работу. Также если муж мало тратит время за работой, то скорей всего, жене необходимо будет работать, чтобы общий доход домохозяйства был достаточным для хорошего проживания.

  Следовательно, будем предполагать, что количество часов, проработанных мужем, негативно влияет на вероятность занятости жены.

- kids5

  Маленькие дети требуют заботы, внимания и воспитания. Эти обязанности чаще всего перекладываются на жену, то есть мать детей. Поэтому при наличии маленьких детей в семье женщины вынуждены много времени тратить на воспитание детей, что ограничивает их от возможности выхода на работу.

  Следовательно, будем предполагать, что количество детей младше 6 лет негативно влияет на вероятность занятости жены.

- unemployment_region

  Общая обстановка в регионе также может влиять на занятость женщин. Если в регионе проживания высокая безработица, то скорей всего там наблюдается плохой инвестиционный климат и малое количество рабочих мест, что даже при наличии желании женщины работать ограничивает ее быть занятой.

  Следовательно, будем предполагать, что безработица в регионе проживания женщины негативно влияет на вероятность занятости жены.

Часть 2. Модель Тобина

Задание №2.1.

Оцените Тобит модель, предварительно записав максимизируемую функцию правдоподобия. Результат представьте в форме таблицы (можно, например, использовать выдачу из stata, R или python).

Тобит модель предполагает оценивание латентной переменной $wage_i^*$, исходя из которой определяется значение наблюдаемой переменной $wage_i$

$$wage_i^* = \beta_0 + \beta_h \cdot husband\_work\_hours_i + \beta_{ed} \cdot education_i + \beta_{ex} \cdot experience_i + \varepsilon_i, \varepsilon_i \sim N(0, \sigma^2)$$

$$wage_i = \begin{cases} wage_i^*, wage_i^* > 0 \\ 0, wage_i^* \leq 0 \end{cases}$$

Исходя из этого, можно легко написать максимизируемую функцию правдоподобия (см. ниже). Где φ - функция плотности стандартного нормального распределения, Ф - функция стандартного нормального распределения, σ - оцениваемое стандартное отклонение $wage_i$ если бы оно не было цензурированно, β - вектор оцениваемых коэффициентов $(\beta_0, \beta_h, \beta_{ed}, \beta_{ex})$'.

$$\mathcal{L} = \prod_{wage_i^* > 0} \left[\frac{1}{\sigma} \varphi \left(\frac{wage_i - (\beta_0 + \beta_h \cdot husband\_work\_hours_i + \beta_{ed} \cdot education_i + \beta_{ex} \cdot experience_i)}{\sigma} \right) \right] \times$$
$$\prod_{wage_i^* \leq 0} \left[1 - \Phi \left(\frac{\beta_0 + \beta_h \cdot husband\_work\_hours_i + \beta_{ed} \cdot education_i + \beta_{ex} \cdot experience_i}{\sigma} \right) \right] \to \max_{\beta, \sigma}$$

Результат оценивания представлен на Рисунке 1.

Из представленной выгрузки из R можно видеть, значения коэффициентов каждой переменной, их стандартные ошибки, z-статистики для проверки гипотезы о значимости коэффицентов (нулевая гипотеза заключается в равенстве коэффициента нулю) и соответствующие p-value. Также можно видеть, что коэффициенты посчитаны благодаря численной оптимизации функции правдоподобия, записанной выше, где 5 - число оцениваемых параметров (4 коэффициента β и стандартное отклонение)

Задание №2.2.

Опишите преимущества Тобит модели над усеченной регрессией. Объясните, в каких случаях можно использовать усеченную регрессию, но не получится использовать Тобит модель: приведите гипотетический (можно использовать фантазию) пример.

Преимущества Тобит модели над усеченной регрессией:

- Асимптотическая нормальность ошибок
- Асимптотическая состоятельность оценок
- Более эффективные оценки
- Не исчезают наблюдения, что позволяет оценить направленность эффекта на полной выборке и не нарушает предпосылку ТГМ

Рис. 1: Результат оценивания Тобит модели в R

```
Call:
crch(formula = wage ~ husband_work_hours + education + experience, data = h, left = tr_left)

Standardized residuals:
    Min      1Q  Median      3Q     Max
-1.6439 -0.0295  0.3033  0.6541  5.7434

Coefficients (location model):
                    Estimate Std. Error z value             Pr(>|z|)
(Intercept)       -8.2630686  1.2264878  -6.737 0.000000000016149063 ***
husband_work_hours -0.0005791  0.0003070  -1.886               0.0593 .
education          0.6816040  0.0806369   8.453 < 0.0000000000000002 ***
experience         0.1840851  0.0224939   8.184 0.000000000000000275 ***

Coefficients (scale model with log link):
            Estimate Std. Error z value            Pr(>|z|)
(Intercept)  1.50214    0.03699   40.62 <0.0000000000000002 ***
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Distribution: gaussian
Log-likelihood: -1470 on 5 Df
Number of iterations in BFGS optimization: 12
```

Усеченную регрессию можно использовать, если цензурированные наблюдения несли неважную информацию и их добавление в модели Тобина, на самом деле, будет вести к не самым эффективным оценкам. То есть в таких случаях, когда усечение необходимо из поставленного вопроса и добавлять пропущенные наблюдения не нужно (они могут не позволить ответить на поставленный вопрос).

Возможным примером может быть исследование факторов роста акций крупных компаний - "голубых фишек". В этом случае не нужно включать мелкие компании - цензурированные наблюдения. Действительно, для акций мелких компаний большее значение может иметь волатильность и различные другие финансовые факторы риска, которые для крупных не так важны, поэтому для анализа поставленного вопроса лучше всего подойдет усеченная модель.

Задание №2.3.

Проинтерпретируйте полученные значения оценок для каждой независимой переменной. Поясните, как полученные результаты соотносятся с высказанными вами ранее предположениями.

По полученным коэффициентам можно оценить предельный эффект переменной на математическое ожидание латентной зависимой переменной (то есть $\frac{\partial \mathbb{E}(wage_i^*)}{\partial x_s} = \hat{\beta}_s$), а также оценить направленность предельного эффекта на математическое ожидание зависимой переменной (то есть $\frac{\partial \mathbb{E}(wage_i)}{\partial x_s} = \hat{\beta}_s \Phi(\frac{x_i'\beta}{\sigma}), 1 \geq \Phi(\frac{x_i'\beta}{\sigma}) \geq 0)$. То есть эффект на зарплату для работающих женщин (для латентной переменной) будет больше по модулю, чем для всей выборки. Исходя из этого приведем интерпретацию полученных коэффициентов.

- husband_work_hours - количество часов проработанных мужем за год имеет отрицательный эффект на уровень зарплаты женщины - увеличение количества часов проработанных мужем на 100 часов в среднем уменьшает почасовую зарплату работающей жены на

0.05 долларов (для общей выборки так явно не можем сказать, для них только можем утверждать про отрицательный эффект).

В целом это подтверждает наше предположение, ведь чем больше работает мужчина, тем меньше нужно работать женщине и зарабатывать соответственно. Хотя коэффициент значим только на 10% уровне значимости (используется z статистика для проверки гипотезы о равенстве коэффициента нулю), что может быть обосновано тем, что мужчина может мало работать в часах, но много зарабатывать или семье оказывает поддержку государство или родители, что тоже уменьшает стимулы женщины работать и больше зарабатывать.

- education - число лет, потраченных женщиной на получение образования влияет положительно на почасовую заработную плату женщины - увеличение числа лет образования на 1 год в среднем увеличивает почасовую зарплату работающей жены на 0.68 долларов (для общей выборки так явно не можем сказать, для них только можем утверждать про положительный эффект).

  Это подтверждает наше предположение о положительном эффекте, ведь большее количество лет обучения ведет к более высокому уровню человеческого капитала, что положительно влияет на заработную плату и востребованность работника. Причем коэффициент по z-статистике значим на любом разумном уровне значимости.

- experience - рабочий стаж женщины также положительно влияет на почасовую заработную плату женщины - увеличение стажа на 1 год в среднем увеличивает почасовую зарплату работающей жены на 0.18 долларов (для общей выборки так явно не можем сказать, для них только можем утверждать про положительный эффект).

  Это подтверждает наше предположение о положительном эффекте, ведь большее число лет в стаже ведет к более высокому уровню профессиональных навыков, что положительно влияет на заработную плату и востребованность работника. Причем коэффициент по z-статистике значим на любом разумном уровне значимости.

Задание №2.4.

Для индивида с произвольными характеристиками укажите (предварительно записав используемые для расчетов формулы):

A) $\mathbb{E}(y^*)$

B) $\mathbb{E}(y)$

C) *Вероятности того, что индивид работает*

Зададим характеристики рассматриваемого индивида в Таблице 1 и рассмотрим каждый пункт отдельно.

Таблица 1: Характеристики рассматриваемого индивида

Переменная	Значение
$husband\_work\_hours_{ind}$	2000
$education_{ind}$	15
$experience_{ind}$	5

A) $\mathbb{E}(y^*)$

Из лекций знаем, что $\mathbb{E}(wage_i^*) = \hat{\beta}_0 + \hat{\beta}_h \cdot husband\_work\_hours_{ind} + \hat{\beta_{ed}} \cdot education_{ind} + \hat{\beta_{ex}} \cdot experience_{ind}$

Подставляя найденные оценки и значения для индивида находим

$$\mathbb{E}(wage_i^*) = -8.2630685960 - 0.0005790926 * 2000 + 0.6816039702 * 15 + 0.1840850533 * 5 = 1.723231$$

То есть для выбранного индивида математическое ожидание почасовой заработной платы при условии работающей женщины будет равна примерно 1.72 доллара

B) $\mathbb{E}(y)$

Из лекций знаем, что

$$\mathbb{E}(wage_i) = (\hat{\beta}_0 + \hat{\beta}_h \cdot husband\_work\_hours_{ind} + \hat{\beta_{ed}} \cdot education_{ind} + \hat{\beta_{ex}} \cdot experience_{ind} + \hat{\sigma} \cdot \lambda_i) \times \times \Phi\left(\frac{\hat{\beta}_0 + \hat{\beta}_h \cdot husband\_work\_hours_{ind} + \hat{\beta_{ed}} \cdot education_{ind} + \hat{\beta_{ex}} \cdot experience_{ind}}{\hat{\sigma}}\right)$$

Где $\lambda_i = \dfrac{\varphi\left(\frac{\hat{\beta}_0+\hat{\beta}_h \cdot husband\_work\_hours_{ind}+\hat{\beta_{ed}} \cdot education_{ind}+\hat{\beta_{ex}} \cdot experience_{ind}}{\hat{\sigma}}\right)}{\Phi\left(\frac{\hat{\beta}_0+\hat{\beta}_h \cdot husband\_work\_hours_{ind}+\hat{\beta_{ed}} \cdot education_{ind}+\hat{\beta_{ex}} \cdot experience_{ind}}{\hat{\sigma}}\right)}$

Подставим найденные оценки и значения для выбранного индивида

$$\lambda_i = \frac{\varphi\left(\frac{-8.2630685960-0.0005790926*2000+0.6816039702*15+0.1840850533*5}{4.49129}\right)}{\Phi\left(\frac{-8.2630685960-0.0005790926*2000+0.6816039702*15+0.1840850533*5}{4.49129}\right)} = \frac{\varphi(0.3836828)}{\Phi(0.3836828)} = 0.5707363$$

$$\mathbb{E}(wage_i) = (-8.2630685960 - 0.0005790926 * 2000 + 0.6816039702 * 15 + 0.1840850533 * 5 + 4.49129 * 0.5707363) \times \Phi(0.3836828) = 2.783672$$

То есть для выбранного индивида математическое ожидание почасовой заработной платы будет равна примерно 2.78 доллара

C) *Вероятности того, что индивид работает*

Из лекций знаем, что

$$P(wage_i > 0) = \Phi\left(\frac{\hat{\beta}_0 + \hat{\beta}_h \cdot husband\_work\_hours_{ind} + \hat{\beta_{ed}} \cdot education_{ind} + \hat{\beta_{ex}} \cdot experience_{ind}}{\hat{\sigma}}\right) = 0.6493932$$

То есть для выбранного индивида вероятности того, что она работает будет равна примерно 0.65.

Задание №2.5.

Для индивида с произвольными характеристиками рассчитайте предельный эффект любой переменной (не дамми), входящей линейно (предварительно записав используемые для расчетов формулы) на:

A) $\mathbb{E}(y^*)$

B) $\mathbb{E}(y)$

C) *Вероятности того, что индивид работает*

Зададим характеристики рассматриваемого индивида в Таблице 1 и рассмотрим каждый пункт отдельно, где будем считать предельный эффект для переменной числа лет, потраченной женщиной на получение образования *education*.

A) $\mathbb{E}(y^*)$

Из лекций знаем, что $\dfrac{\partial \mathbb{E}(wage_i^*)}{\partial education} = \hat{\beta_{ed}} = 0.681604$

То есть увеличение числа лет образования на 1 год для выбранного индивида может увеличить почасовую зарплату работающей женщины (то есть при условии что она работает) примерно на 0.68 долларов

B) $\mathbb{E}(y)$

Из лекций знаем, что

$\dfrac{\partial \mathbb{E}(wage_i)}{\partial education} = \hat{\beta_{ed}} \times \Phi\left(\frac{\hat{\beta_0}+\hat{\beta_h}\cdot husband\_work\_hours_{ind}+\hat{\beta_{ed}}\times education_{ind}+\hat{\beta_{ex}}\cdot experience_{ind}}{\hat{\sigma}}\right) = 0.442629$

То есть увеличение числа лет образования на 1 год для выбранного индивида может увеличить почасовую зарплату примерно на 0.44 доллара (рассчитывая по всей выборке вне зависимости работала или нет).

C) *Вероятности того, что индивид работает*

Из лекций знаем, что

$\dfrac{\partial P(wage_i > 0)}{\partial educ} = \frac{\hat{\beta_{ed}}}{\hat{\sigma}} \times \varphi\left(\frac{\hat{\beta_0}+\hat{\beta_h}\cdot husband\_work\_hours_{ind}+\hat{\beta_{ed}}\cdot education_{ind}+\hat{\beta_{ex}}\cdot experience_{ind}}{\hat{\sigma}}\right) = 0.05624763$

То есть увеличение числа лет образования на 1 год для выбранного индивида может увеличить вероятность стать занятым на 5%.

Задание №2.6.*

Добавьте в модель нелинейный эффект (например, квадрат). Повторите предыдущий пункт для переменной, имеющей нелинейный эффект.

Исследуем квадратичное влияние на почасовую зарплату женщин у количества лет потраченное на получение образования. По нескольким причинам будем предполагать обнаружить перевернутую U-образную зависимость. Причинами для такого возможного наличия следующие: при большом количестве лет обучения после определенного уровня происходит переобучение женщин и их зарплата будет только уменьшаться с увеличением количества лет обучения, так как на рынке труда не нужны переквалифицированные люди с множеством навыков; также возможно женщины с большим количеством лет получения образования собираются идти в определенные профессии, которые на самом деле менее оплачиваемые (например, в науку). Поэтому новая модель будет задано следующим уравнениями:

$$wage_i^* = \beta_0 + \beta_h \cdot husband\_work\_hours_i + \beta_{ed} \cdot education_i + \beta_{ex} \cdot experience_i + \beta_{ed2} \cdot education_i^2 + \varepsilon_i$$

$$wage_i = \begin{cases} wage_i^*, wage_i^* > 0 \\ 0, wage_i^* \leq 0 \end{cases}$$

Результаты оценивания такой регрессии представлены на Рисунке 2

Рис. 2: Результат оценивания Тобит модели в R с нелинейным вхождением лет образования

```
Call:
crch(formula = wage ~ husband_work_hours + education + experience + I(education^2), data = h, left = tr_left)

Standardized residuals:
    Min      1Q  Median      3Q     Max
-1.7143 -0.0011  0.3059  0.6541  5.8154

Coefficients (location model):
                     Estimate Std. Error z value             Pr(>|z|)
(Intercept)        -1.7704334  3.4387867  -0.515                0.6067
husband_work_hours -0.0005386  0.0003057  -1.762                0.0781 .
education          -0.3955172  0.5432587  -0.728                0.4666
experience          0.1833193  0.0223518   8.202 0.000000000000000237 ***
I(education^2)      0.0427590  0.0214038   1.998                0.0457 *

Coefficients (scale model with log link):
            Estimate Std. Error z value            Pr(>|z|)
(Intercept)  1.49653    0.03701   40.44 <0.0000000000000002 ***
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Distribution: gaussian
Log-likelihood: -1468 on 6 Df
Number of iterations in BFGS optimization: 24
```

Из представленной выгрузки из R можно видеть, значения коэффициентов каждой переменной, их стандартные ошибки, z-статистики для проверки гипотезы о значимости коэффицентов (нулевая гипотеза заключается в равенстве коэффициента нулю) и соответствующие p-value. Также можно видеть, что коэффициенты посчитаны благодаря численной оптимизации функции правдоподобия, записанной выше, где 6 - число оцениваемых параметров (5 коэффициента β и стандартное отклонение). Также видно, что коэффициент перед количеством часов проработанных мужем также значим только на 10% уровне значимости, число лет рабочего стажа значим на любом разумном уровне значимости, а число лет потраченное на получение образования при линейном вхождении не значим, а при квадратичном значим на 5% уровне значимости, что частично подтверждает наше предположение, но необходимо посчитать предельные эффекты.

Посчитаем предельные эффекты для выбранного индивида, чьи характеристики представлены в Таблице 1.

A) $\mathbb{E}(y^*)$

Из лекций знаем, что

$$\mathbb{E}(wage_i^*) = \hat{\beta}_0 + \hat{\beta}_h \cdot husband\_work\_hours_i + \hat{\beta_{ed}} \cdot education_i + \hat{\beta_{ed2}} \cdot education_i^2 + \hat{\beta_{ex}} \cdot experience_i$$

Так как математическое ожидание не зависит от форм вхождения независимых переменных. Отсюда легко найдем предельный эффект:

$$\frac{\partial \mathbb{E}(wage_i^*)}{\partial education} = \hat{\beta_{ed}} + 2 * \hat{\beta_{ed2}} * education_i = 0.8872537$$

То есть увеличение числа лет образования на 1 год для выбранного индивида может увеличить почасовую зарплату работающей женщины примерно на 0.88 долларов

B) $\mathbb{E}(y)$

Из лекций знаем, что $E(wage_i)$ имеет такую же формулу, так как математическое ожидание не зависит от форм вхождения независимых переменных.

$\mathbb{E}(wage_i) = (\hat{\beta}_0 + \hat{\beta}_h \cdot husband\_work\_hours_i + \hat{\beta_{ed}} \cdot education_i + \hat{\beta_{ed2}} \cdot education_i^2 + \hat{\beta_{ex}} \cdot experience_i + \hat{\sigma} \cdot \lambda_i) \times \Phi\left(\frac{\hat{\beta}_0+\hat{\beta}_h \cdot husband\_work\_hours_i+\hat{\beta_{ed}} \cdot education_i+\hat{\beta_{ed2}} \cdot education_i^2+\hat{\beta_{ex}} \cdot experience_i}{\hat{\sigma}}\right)$

Где $\lambda_i = \dfrac{\varphi\left(\frac{\hat{\beta}_0+\hat{\beta}_h \cdot husband\_work\_hours_i+\hat{\beta_{ed}} \cdot education_i+\hat{\beta_{ed2}} \cdot education_i^2+\hat{\beta_{ex}} \cdot experience_i}{\hat{\sigma}}\right)}{\Phi\left(\frac{\hat{\beta}_0+\hat{\beta}_h \cdot husband\_work\_hours_i+\hat{\beta_{ed}} \cdot education_i+\hat{\beta_{ed2}} \cdot education_i^2+\hat{\beta_{ex}} \cdot experience_i}{\hat{\sigma}}\right)}$

Путем взятия производных и подсчета можно получить, что

$\dfrac{\partial \mathbb{E}(wage_i)}{\partial education} = (\hat{\beta_{ed}} + 2 * \hat{\beta_{ed2}} * education_i) \times$

$\times \Phi\left(\frac{\hat{\beta}_0+\hat{\beta}_h \cdot husband\_work\_hours_i+\hat{\beta_{ed}} \cdot education_i+\hat{\beta_{ed2}} \cdot education_i^2+\hat{\beta_{ex}} \cdot experience_i}{\hat{\sigma}}\right) = 0.5793682$

То есть увеличение числа лет образования на 1 год для выбранного индивида может увеличить почасовую зарплату примерно на 0.58 доллара (рассчитывая по всей выборке вне зависимости работала или нет).

C) *Вероятности того, что индивид работает*

Из лекций знаем, что

$P(wage_i > 0) = \Phi\left(\frac{\hat{\beta}_0+\hat{\beta}_h \cdot husband\_work\_hours_i+\hat{\beta_{ed}} \cdot education_i+\hat{\beta_{ed2}} \cdot education_i^2+\hat{\beta_{ex}} \cdot experience_i}{\hat{\sigma}}\right)$

Путем взятия производных и подсчета можно получить, что

$$\frac{\partial P(wage_i > 0)}{\partial education} = \frac{\hat{\beta_{ed}} + 2 * \hat{\beta_{ed2}} * education_i}{\hat{\sigma}} \times$$
$$\times \varphi\left(\frac{\hat{\beta}_0 + \hat{\beta}_h \cdot husband\_work\_hours_i + \hat{\beta_{ed}} \cdot education_i + \hat{\beta_{ed2}} \cdot education_i^2 + \hat{\beta_{ex}} \cdot experience_i}{\hat{\sigma}}\right) =$$
$$= 0.07335294$$

То есть увеличение числа лет образования на 1 год для выбранного индивида может увеличить вероятность стать занятым на 7%.

Задание №2.7.**

При помощи LR теста проверьте гипотезу о гомоскедастичности в Тобит модели, предварительно формально записав предполагаемые нулевой гипотезой ограничения на параметры, асимптотическое распределение тестовой статистики (при верной нулевой гипотезе) и максимизируемую в гетероскедастичной Тобит модели функцию правдоподобия. При этом уравнение дисперсии должно включать по крайней мере одну переменную, не входящую в основное

уравнение. Укажите негативные последствия, к которым может приводить отсутствие учета гетероскедастичности при условии ее наличия в Тобит модели.

В гетероскедастичной Тобит модели будем предполагать, что дисперсия ошибки зависит от возраста (чем больше возраст, тем больше разброс в зарплатах, кто-то становится более успешным, кто-то менее успешным), а также будет зависеть от рабочего стажа (аналогично возрасту с более высоким стажем увеличивается разброс в зарплатах). Причем учитывая, что при оценивании Тобит модели находится логарифм стандартной ошибки, то будем предполагать такую зависимость $log(\sigma_i) = \gamma_0 + \gamma_a * age_i + \gamma_e * experience_i + u_i$.

В гетероскедастичной модели Тобита в функции правдоподобия будет также включена оценка параметров для ошибок, а именно

$$\mathcal{L} = \prod_{wage_i^* > 0} \left[\frac{1}{exp(\gamma_0 + \gamma_a * age_i + \gamma_e * experience_i)} \times \varphi\left(\frac{wage_i - (\beta_0 + \beta_h \cdot husband\_work\_hours_i + \beta_{ed} \cdot education_i + \beta_{ex} \cdot experience_i)}{\gamma_0 + \gamma_a * age_i + \gamma_e * experience_i} \right) \right] \times \prod_{wage_i^* \leq 0} \left[1 - \Phi\left(\frac{\beta_0 + \beta_h \cdot husband\_work\_hours_i + \beta_{ed} \cdot education_i + \beta_{ex} \cdot experience_i}{\gamma_0 + \gamma_a * age_i + \gamma_e * experience_i} \right) \right] \to \max_{\beta, \gamma}$$

Где φ - функция плотности стандартного нормального распределения, Φ - функция стандартного нормального распределения, σ - оцениваемое стандартное отклонение $wage_i$ если бы оно не было цензурированно, β - вектор оцениваемых коэффициентов $(\beta_0, \beta_h, \beta_{ed}, \beta_{ex})$', γ - вектор оцениваемых коэффициентов $(\gamma_0, \gamma_a, \gamma_e)$'. То есть в этой модели оцениваемых параметров 7.

Результаты оценивания представлены на Рисунке 3.

Рис. 3: Результат оценивания гетероскедастичной Тобит модели в R

```
Call:
crch(formula = wage ~ husband_work_hours + education + experience | age + experience, data = h, link.scale = "log",
    left = tr_left)

Standardized residuals:
    Min      1Q  Median      3Q     Max
-2.5551 -0.0208  0.2744  0.6437  5.1138

Coefficients (location model):
                    Estimate Std. Error z value             Pr(>|z|)
(Intercept)       -8.5581320  1.1987051  -7.139    0.000000000000937 ***
husband_work_hours -0.0005144  0.0002931  -1.755               0.0792 .
education          0.7052441  0.0771014   9.147 < 0.0000000000000002 ***
experience         0.1718964  0.0193901   8.865 < 0.0000000000000002 ***

Coefficients (scale model with log link):
             Estimate Std. Error z value        Pr(>|z|)
(Intercept)  1.351877   0.192539   7.021 0.0000000000022 ***
age          0.010065   0.004875   2.065           0.039 *
experience  -0.024980   0.004374  -5.711 0.0000000112327 ***
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Distribution: gaussian
Log-likelihood: -1454 on 7 Df
Number of iterations in BFGS optimization: 24
```

В LR тесте на проверку гомоскедастичности нулевая гипотеза будет заключаться в том, что $\gamma_a = \gamma_e = 0$. Ограниченная модель будет модель Тобита без учета возможной гетероскедастичности, а полная модель будет модель Тобита с учетом возможной гетероскедастичности.

$$\begin{cases} H0 : \gamma_a = \gamma_e = 0 \\ Ha : \gamma_a^2 + \gamma_e^2 \neq 0 \end{cases}$$

При этом как обычно тестовая статистика при верной нулевой гипотезе будет равна $2 * (LR_F - LR_R) \sim \chi^2_{(m-n)}$, где LR - значения функций правдоподобий в полной и ограниченной моделях, m и n - количество параметров, оцениваемых в моделях, в нашем случае они равны 7 и 5, то есть их разница будет 2.

Таким образом получаем, что

$LR_{statistic} = 2 * (LR_F - LR_R) = 2 * (-1454.415 - (-1470.356)) = 31.88393 \sim \chi^2_2$

Откуда получаем, что p-value = 0.0000001192596, то есть на лубом разумном уровне значимости нулевая гипотеза отвергается, а значит существует гетероскедастичность в модели.

Не учет гетероскедастичности может привести к таким негативным последствиям

- Неэффективность оценок
- Возможная потеря состоятельности оценок
- Искажение в оценивании предельных эффектов (особенно для тех факторов, которые еще влияют на гетероскедастичность)

Задание №2.8.***

Для индивида с произвольными характеристиками в Тобит модели с гетероскедастичной случайной ошибкой рассчитайте (предварительно записав соответствующую формулу) предельный эффект переменной (не дамми), входящей и в основное уравнение, и в уравнение дисперсии, на:

A) $\mathbb{E}(y^*)$

B) $\mathbb{E}(y)$

C) *Вероятности того, что индивид работает*

Посчитаем предельные эффект для переменной *experience*, входящей в оба уравнения по пунктам для индивида, чьи характеристики представлены в Таблице 2

Таблица 2: Характеристики рассматриваемого индивида

Переменная	Значение
$husband\_work\_hours_{ind}$	2000
$education_{ind}$	15
$experience_{ind}$	5
age_{ind}	45

A) $\mathbb{E}(y^*)$

Так как в гетероскедастичной модели Тобита математическое ожидание не изменяется и является просто линейным индексом с оцененными параметрами β, то и предельный эффект тоже будет равен коэффициенту $\hat{\beta_{ex}}$.

Таким образом

$$\frac{\partial \mathbb{E}(wage_i^*)}{\partial experience} = \hat{\beta_{ex}} = 0.1718964$$

То есть увеличение числа лет рабочего стажа на 1 год для выбранного индивида может увеличить почасовую зарплату работающей женщины примерно на 0.17 долларов

B) $\mathbb{E}(y)$

Зная из лекций, что $\mathbb{E}(wage_i) = x_i'\hat{\beta} * \Phi(\frac{x_i'\hat{\beta}}{\hat{\sigma}}) + \hat{\sigma} * \varphi(\frac{x_i'\hat{\beta}}{\hat{\sigma}}) = x_i'\hat{\beta} * \Phi(\frac{x_i'\hat{\beta}}{exp(z_i'\hat{\gamma})}) + exp(z_i'\hat{\gamma}) * \varphi(\frac{x_i'\hat{\beta}}{exp(z_i'\hat{\gamma})})$

Где Φ - функция распределения у стандартного нормального распределения, φ - функция плотности стандартного нормального распределения, $\hat{\beta}$ - вектор оцененных коэффициентов $(\hat{\beta}_0, \hat{\beta}_h, \hat{\beta_{ed}}, \hat{\beta_{ex}})$', x_i' - вектор значений переменных (1, $husband\_work\_hours_i$, $education_i$, $experience_i$), $\hat{\gamma}$ - вектор оцененных коэффициентов $(\hat{\gamma}_0, \hat{\gamma}_a, \hat{\gamma}_e)$', z_i' - вектор значений переменных в уравнении дисперсии (1, age_i, $experience_i$).

То путем взятия производных и вычислений нетрудно получить, что

$$\frac{\partial \mathbb{E}(wage_i)}{\partial experience} = \hat{\beta_{ed}} \times \Phi\left(\frac{\hat{\beta}_0 + \hat{\beta}_h \cdot husband\_work\_hours_{ind} + \hat{\beta_{ed}} \times education_{ind} + \hat{\beta_{ex}} \cdot experience_{ind}}{\hat{\sigma}}\right) + \gamma_e * \hat{\sigma} * \varphi\left(\frac{\hat{\beta}_0 + \hat{\beta}_h \cdot husband\_work\_hours_{ind} + \hat{\beta_{ed}} \times education_{ind} + \hat{\beta_{ex}} \cdot experience_{ind}}{\hat{\sigma}}\right) = 0.05877214$$

То есть увеличение числа лет стажа на 1 год для выбранного индивида может увеличить почасовую зарплату примерно на 0.06 доллара (рассчитывая по всей выборке вне зависимости работала или нет).

C) *Вероятности того, что индивид работает*

Из лекций знаем, что

$$P(wage_i > 0) = \Phi\left(\frac{\hat{\beta}_0 + \hat{\beta}_h \cdot husband\_work\_hours_{ind} + \hat{\beta_{ed}} \cdot education_{ind} + \hat{\beta_{ex}} \cdot experience_{ind}}{\hat{\sigma}}\right)$$

Путем взятия производных нетрудно получить, что

$$\frac{\partial P(wage_i > 0)}{\partial experience} = \varphi\left(\frac{\hat{\beta}_0 + \hat{\beta}_h \cdot husband\_work\_hours_{ind} + \hat{\beta_{ed}} \cdot education_{ind} + \hat{\beta_{ex}} \cdot experience_{ind}}{\hat{\sigma}}\right) \times$$
$$\times\left(\frac{\hat{\beta_{ex}}}{\hat{\sigma}} - (\hat{\beta}_0 + \hat{\beta}_h \cdot husband\_work\_hours_{ind} + \hat{\beta_{ed}} \cdot education_{ind} + \hat{\beta_{ex}} \cdot experience_{ind}) * \frac{\hat{\gamma_e}}{\hat{\sigma}}\right) =$$

$$= 0.01528326$$

То есть увеличение числа лет рабочего стажа на 1 год для выбранного индивида может увеличить вероятность стать занятым на 1.5%.

Часть 3. Модель Хекмана

Задание №3.1.

Оцените модель Хекмана с помощью метода максимального правдоподобия, предварительно записав максимизируемую функцию правдоподобия и указав независимые переменные в уравнении занятости, которое должно иметь по крайней мере одну переменную, не входящую в уравнение зарплаты. Результат представьте в форме таблицы (можно, например, использовать выдачу из stata, R или python).

В модели Хекмана задаются два уравнения с отдельными латентными зависимыми переменными: уравнения участия 1 и уравнения интенсивности 2. Причем считается, что ошибки имеют совместное нормальное распределение с коэффициентом корреляции ρ.

Исходя из уравнения участия (дамми на занятость) находится $employment\_dummy_i$, по которому определяется наблюдаем ли мы зарплату или нет, и какую зарплату наблюдаем.

$$employment\_dummy_i^* = \gamma_0 + \gamma_h \cdot husband\_work\_hours_i + \gamma_k \cdot kids5_i + \gamma_u \cdot unemployment\_region_i + u_i \tag{1}$$

$$wage_i^* = \beta_0 + \beta_h \cdot husband\_work\_hours_i + \beta_{ed} \cdot education_i + \beta_{ex} \cdot experience_i + \varepsilon_i \tag{2}$$

$$\begin{pmatrix} \varepsilon_i \\ u_i \end{pmatrix} \sim N\left[\begin{pmatrix} 0 \\ 0 \end{pmatrix}, \quad \begin{pmatrix} \sigma^2 & \rho\sigma \\ \rho\sigma & 1 \end{pmatrix}\right]$$

$$employment\_dummy_i = \begin{cases} 1, employment\_dummy_i^* > 0 \\ 0, employment\_dummy_i^* \leq 0 \end{cases}$$

$$wage_i = \begin{cases} wage_i^*, employment\_dummy_i = 1 \\ \text{не наблюдаем}, employment\_dummy_i = 0 \end{cases}$$

Данную модель можно оценить методом максимального правдоподобия, разделив функции правдоподобия для тех, кто устроен на работу, и для тех кто не работает. Для работающих женщин функция правдоподобия будет строится на основе двух уравнений (используя совместное нормальное распределение ошибок), а для неработающих только на основе уравнения участия. То есть максимизируемая функция правдоподобия будет равна 3.

Оцениваемыми параметрами являются векторы коэффициентов β, γ из уравнений 2, 1, а также дисперсия ошибок ε_i равная σ, а также коэффициент корреляции между ошибками уравнений (в уравнении x_i' - вектор значений переменных (1, $husband\_work\_hours_i$, $education_i$, $experience_i$), z_i- вектор значений переменных (1, $husband\_work\_hours_i$, $kids5_i$, $unemployment\_region_i$).

$$\mathcal{L} = \prod_{employment\_dummy_i=1} \left\{ \left[\frac{1}{\sigma} \varphi \left(\frac{wage_i - (x_i' * \beta)}{\sigma} \right) \right] \times \left[1 - \Phi \left(\frac{-(z_i'\gamma + \rho \cdot \frac{wage_i - (x_i' * \beta)}{\sigma})}{\sqrt{1-\rho^2}} \right) \right] \right\} \times \prod_{employment\_dummy_i=0} (1 - \Phi(z_i'\gamma)) \to \max_{\beta,\gamma,\sigma,\rho} \tag{3}$$

Результаты оценивания такой регрессии представлены на Рисунке 4.

Из представленной выгрузки из R можно видеть, значения коэффициентов каждой переменной в каждом из 2 уравнений, их стандартные ошибки, t-статистики для проверки гипотезы

Рис. 4: Результат оценивания модели Хекмана в R, основанный на ММП

```
Tobit 2 model (sample selection model)
Maximum Likelihood estimation
Newton-Raphson maximisation, 3 iterations
Return code 8: successive function values within relative tolerance limit (reltol)
Log-Likelihood: -1585.99
753 observations (325 censored and 428 observed)
10 free parameters (df = 743)
Probit selection equation:
                      Estimate  Std. Error t value     Pr(>|t|)
(Intercept)          0.78693877  0.24473844   3.215      0.00136 **
husband_work_hours  -0.00014902  0.00007928  -1.880      0.06055 .
kids5               -0.54031638  0.09468402  -5.707 0.0000000167 ***
unemployment_region -0.01727969  0.01514913  -1.141      0.25439
Outcome equation:
                     Estimate Std. Error t value           Pr(>|t|)
(Intercept)        -1.6927678  1.0593581  -1.598              0.110
husband_work_hours -0.0003383  0.0002624  -1.289              0.198
education           0.5059552  0.0661087   7.653 0.0000000000000609 ***
experience          0.0222400  0.0187629   1.185              0.236
   Error terms:
      Estimate Std. Error t value            Pr(>|t|)
sigma   3.0953     0.1065  29.056 <0.0000000000000002 ***
rho    -0.0335     0.2043  -0.164                0.87
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1
```

о значимости коэффициентов (нулевая гипотеза заключается в равенстве коэффициента нулю) и соответствующие p-value, также оценены $\sigma = 3.09, \rho = -0.0335$ и их тестовые статистики, по которым понятно, что коэффициент корреляции не значим. . Более того, можно видеть, что коэффициенты посчитаны благодаря численной оптимизации функции правдоподобия, записанной выше. Можно заметить и число наблюдений - 753 женщины, из которых 325 не работают и цензурированы.

Относительно изложенных предположений о направленности эффектов в каждом из уравнений, то все коэффициенты соответствуют предположениям (они же отображают направленность предельных эффектов на латентные переменные), но малая их часть значима в каждом из уравнений. В уравнении участия значимы $husband\_work\_hours$ - на 10% уровне значимости, а kids5 на любом разумном уровне значимости. В уравнении интенсивности значимо только количество лет для получения образования на любом разумном уровне значимости.

Задание №3.2.

Опишите отличия модели Хекмана от модели Тобина.

Можно назвать следующие отличия модели Хекмана от модели Тобина:

- Не включение цензурированных наблюдений при оценивании главного уравнения, что позволяет уйти от возможных аутлаеров около нуля в модели Тобина (да и в целом в модели Хекмана логика оценивания более верная, так как в ней в основном уравнении не учитываются те, которые не прошли уравнение участие)
- Отдельный учет условия участия в другом уравнении, а не совместно, что позволяет более верно оценить эффекты по выборке (может быть так, что факторы влияют на интенсивность при участии положительно, а при участии нейтрально или отрицательно)

- Возможность оценивания других факторов на принятие решения о участии, которые могут не влиять на интенсивность, что позволяет более полно оценивать вероятность участия

Задание №3.3.

Воспользуйтесь методом Хекмана, основанным на двухшаговой процедуре и сравните оценки, с полученными с использованием метода Хекмана, основанном на методе максимального правдоподобия. Опишите относительные преимущества и недостатки обоих методов.

Двухшаговая процедура метода Хекмана предполагает на первом шаге оценивания уравнения участия по пробит-модели, из которого находим коэффициенты γ, а также считаем λ_i по формуле аналогичной в модели Тобина, а на втором шаге благодаря МНК находим коэффициенты β в уравнении участия с учетом λ_i. То есть для выбранных переменных схематично можно записать так:

1. Probit

$$P(employment\_dummy_i = 1) = \gamma_0 + \gamma_h \cdot husband\_work\_hours_i + \gamma_k \cdot kids5_i + \\ + \gamma_u \cdot unemployment\_region_i + u_i$$

   Отсюда находим оценки для $\gamma_h, \gamma_k, \gamma_u$, а также

$$\hat{\lambda_i} = \frac{\varphi(\hat{\gamma_0} + \hat{\gamma_h} \cdot husband\_work\_hours_i + \hat{\gamma_k} \cdot kids5_i + \hat{\gamma_u} \cdot unemployment\_region_i)}{\Phi(\hat{\gamma_0} + \hat{\gamma_h} \cdot husband\_work\_hours_i + \hat{\gamma_k} \cdot kids5_i + \hat{\gamma_u} \cdot unemployment\_region_i)}$$

2. МНК

$$wage_i = \beta_h \cdot husband\_work\_hours_i + \beta_{ed} \cdot education_i + \beta_{ex} \cdot experience_i + \beta_\lambda \cdot \hat{\lambda}_i + \varepsilon_i$$

   Отсюда находим оценки для $\beta_h, \beta_{ed}, \beta ex$

Результаты оценивания методом Хекмана по двухшаговой процедуре представлены на Рисунке 5.

Рис. 5: Результат оценивания модели Хекмана в R, основанный на двухшаговой процедуре

```
Tobit 2 model (sample selection model)
2-step Heckman / heckit estimation
753 observations (325 censored and 428 observed)
11 free parameters (df = 743)
Probit selection equation:
                      Estimate  Std. Error t value     Pr(>|t|)
(Intercept)          0.78551191  0.24462361   3.211      0.00138 **
husband_work_hours  -0.00014900  0.00007928  -1.879      0.06060 .
kids5               -0.54030954  0.09468862  -5.706 0.0000000167 ***
unemployment_region -0.01711845  0.01512272  -1.132      0.25801
Outcome equation:
                     Estimate Std. Error t value           Pr(>|t|)
(Intercept)       -1.6326951  1.1291957  -1.446              0.149
husband_work_hours -0.0003292  0.0002692  -1.223              0.222
education          0.5072878  0.0666688   7.609 0.0000000000000838 ***
experience         0.0218178  0.0189610   1.151              0.250
Multiple R-Squared:0.1243,      Adjusted R-Squared:0.116
   Error terms:
              Estimate Std. Error t value Pr(>|t|)
invMillsRatio -0.24204    1.09894   -0.22    0.826
sigma          3.09962         NA      NA       NA
rho           -0.07809         NA      NA       NA
```

Из представленной выгрузки из R можно видеть, значения коэффициентов каждой переменной в каждом из 2 уравнений, их стандартные ошибки, t-статистики для проверки гипотезы о значимости коэффициентов (нулевая гипотеза заключается в равенстве коэффициента нулю) и соответствующие p-value, также оценены $\sigma = 3.09, \rho = -0.078$, а также представлен p-value для β_λ. Более того, можно видеть, что коэффициенты посчитаны по 2 шаговой процедуре, в которой число наблюдений - 753 женщины, из которых 325 не работают и цензурированы.

Относительно изложенных предположений о направленности эффектов в каждом из уравнений, то все коэффициенты соответствуют предположениям (они же отображают направленность предельных эффектов на латентные переменные), но малая их часть значима в каждом из уравнений. В уравнении участия значимы $hushrs$ - на 10% уровне значимости, а kids5 на любом разумном уровне значимости. В уравнении интенсивности значимо только количество лет для получения образования на любом разумном уровне значимости. Также для outcome уравнения посчитаны $AdjustedR^2 = 0.116$, что говорит о не сильной объясняющей силе (это вызвано небольшим количеством объясняющих переменных, которые практически незначимы).

Таблица 3: Сравнение оценок, полученные методом Хекмана на основе ММП и 2-шаговой процедуре

Параметр в уравнении участия	Heckman ММП	Heckman 2-шаговая процедура
Constant	-1.6927677647	-1.6326951058
husband_work_hours	-0.0003382973	-0.0003292103
education	0.5059552311	0.5072877547
experience	0.0222400455	0.0218177973
σ	3.095277	3.09962
ρ	-0.03350136	-0.07808857
p-value для ρ	0.87	0.82

Сравнивая оценки в двух моделях, представленные в Таблице 3, можно отчетливо видеть, что оценки практически идентичны, кроме коэффициента корреляции, о которой пойдет речь в следующем задании. Такая схожесть говорит о том, что оба метода могут быть применены, но у каждой из них есть свои недостатки и преимущества.

Метод Хекмана на основе ММП

- Преимущества
  - Асимптотически состоятельные
  - Эффективней оценки
  - Меньше параметров оценивается
- Недостатки
  - Сильно зависит от предполагаемого распределения ошибок
  - Сложно для понимания и быстрого оценивания
  - Возможен не единственный максимум

Метод Хекмана на основе 2 шаговой процедуры

- Преимущества
  - Проще для оценивания
  - Устойчивей к нарушению о предположении распределений ошибок

  - Лучше при использовании с несколькими условиями участия
  - Всегда есть единственный максимум
  - Позволяет оценить параметр сдвига из-за первоначального выбора участия
- Недостатки
  - Гетероскедастичность ошибок в уравнении интенсивности
  - Менее эффективные оценки

Что из них выбирать? Лучше брать на основе ММП с проверкой с 2 шаговой процедурой и просто сравнивать коэффициенты и их значимость, так как Хекман на основе ММП позволяет получить более хорошие оценки, но возможен не единственный максимум функции правдоподобия, поэтому лучше проверить эти оценки с 2 шаговой процедурой и использовать на основе ММП.

Задание №3.4.

Проинтерпретируйте значимость и значение оценки корреляции между случайными ошибками в обоих оцененных моделях. Укажите, можно ли было бы обойтись оцениванием обычной МНК модели.

Коэффициент ρ показывает корреляцию между ошибками из уравнений участия и интенсивности. Как видно из Таблицы 3 коэффициенты корреляции для метода ММП и двухшаговой процедуры сильно отличаются, хотя оба отрицательные и близки к нулю. Это значение говорит о том, что возможно существует какой-то общий фактор, который влияет по-разному на дамми занятости и зарплату. Но обращая внимание на p-value в Таблице 3 видно, что коэффициенты незначимы, то есть нулевая гипотеза о равенстве корреляции нулю не отвергается. Следовательно, при таких предположения мы не наблюдаем неслучайный отбор. А значит, можно было бы обойтись использованием МНК, так как нет обязанности оценивать участие в модели, которое нам обеспечивает состоятельные оценки.

Давайте оценим МНК, посмотрим на коэффициенты и сравним с моделями Хекмана в Таблице 4. Легко можно заметить, что коэффициенты в целом достаточно близки (особенно для переменной husband_work_hours), что подтверждает выше заключение о том, что можно было бы использовать МНК.

Таблица 4: Сравнение оценок, полученные методом Хекмана на основе ММП, 2-шаговой процедуре, МНК

Параметр в уравнении участия	Heckman ММП	Heckman 2-шаговая процедура	МНК
Constant	-1.6927677647	-1.6326951058	-3.2046182832
husband_work_hours	-0.0003382973	-0.0003292103	-0.0003374964
education	0.5059552311	0.5072877547	0.4384966729
experience	0.0222400455	0.0218177973	0.0899870726

Задание №3.5.

В любой из двух оцененных в данном разделе моделей для индивида с произвольными характеристиками рассчитайте (предварительно записав формулу):

А) $\mathbb{E}(y^*|z=1)$ и $\mathbb{E}(y^*|z=0)$

B) *предельный эффект любой переменной (не дамми), входящей линейно и в основное уравнение, и в уравнение занятости, на* $\mathbb{E}(y^*|z=1)$ *и* $\mathbb{E}(y^*|z=0)$

Будем использовать модель Хекмана на основе ММП, так как они дают более состоятельные оценки. Рассчитаем необходимые величины для индивида, чьи характеристики указаны в Таблице 5

Таблица 5: Характеристики рассматриваемого индивида

Переменная	Значение
$husband\_work\_hours_{ind}$	2000
$education_{ind}$	15
$experience_{ind}$	5
$unemployment\_region_{ind}$	5
$kids5_{ind}$	1
age_{ind}	35

A) $\mathbb{E}(y^*|z=1)$ и $\mathbb{E}(y^*|z=0)$

Из лекций знаем, что

$\mathbb{E}(wage_i^*|employment\_dummy_i=1) = \hat{\beta}_0 + \hat{\beta}_h \cdot husband\_work\_hours_i + \hat{\beta_{ed}} \cdot education_i + \hat{\beta_{ex}} \cdot experience_i + \hat{\rho} * \hat{\sigma} * \hat{\lambda}_i$

Где $\hat{\lambda}_i = \dfrac{\varphi(\hat{\gamma}_0 + \hat{\gamma}_h \cdot husband\_work\_hours_i + \hat{\gamma}_k \cdot kids5_i + \hat{\gamma}_u \cdot unemployment\_region_i)}{\Phi(\hat{\gamma}_0 + \hat{\gamma}_h \cdot husband\_work\_hours_i + \hat{\gamma}_k \cdot kids5_i + \hat{\gamma}_u \cdot unemployment\_region_i)}$

Аналогично знаем, что

$\mathbb{E}(wage_i^*|employment\_dummy_i=0) = \hat{\beta}_0 + \hat{\beta}_h \cdot husband\_work\_hours_i + \hat{\beta_{ed}} \cdot education_i + \hat{\beta_{ex}} \cdot experience_i - \hat{\rho} * \hat{\sigma} * \tilde{\lambda}_i$

Где $\tilde{\lambda}_i = \dfrac{\varphi(\hat{\gamma}_0 + \hat{\gamma}_h \cdot husband\_work\_hours_i + \hat{\gamma}_k \cdot kids5_i + \hat{\gamma}_u \cdot unemployment\_region_i)}{\Phi(-(\hat{\gamma}_0 + \hat{\gamma}_h \cdot husband\_work\_hours_i + \hat{\gamma}_k \cdot kids5_i + \hat{\gamma}_u \cdot unemployment\_region_i))}$

Можно посчитать руками, но в R еще встроен пакет для быстрого подсчета, поэтому воспользуемся им и найдем искомые значения:

$\mathbb{E}(wage_i^*|employment\_dummy_i=1) = 5.233082$

$\mathbb{E}(wage_i^*|employment\_dummy_i=0) = 5.399693$

То есть для этого индивида в среднем зарплата будет равна 5.23 долларов в час при условии, что она будет работать, а при условии, что она не работает ее возможная альтернативная зарплата в среднем могла бы быть равна 5.4 долларов в час.

B) *предельный эффект любой переменной (не дамми), входящей линейно и в основное уравнение, и в уравнение занятости, на* $\mathbb{E}(y^*|z=1)$ *и* $\mathbb{E}(y^*|z=0)$

Так как у меня в оба уравнения входит только одна непрерывная переменная

$husband\_work\_hours$, то будем рассчитывать предельные эффекты для нее.

Чтобы рассчитать предельные эффекты достаточно взять производные по формулам выше и получить, что (эта формула выводилась и на лекции)

$$\frac{\partial \mathbb{E}(wage_i^*|employment\_dummy_i=1)}{\partial husband\_work\_hours} = \hat{\beta}_h - \hat{\rho} * \hat{\sigma} * (z_i'\hat{\gamma} * \hat{\lambda}_i + \hat{\lambda_i}^2) * \hat{\gamma}_h = -0.0003488455$$

То есть для этого индивида при увеличении количества рабочих часов у мужа на 100 зарплата при условии, что она будет работать, упадет в среднем на 0.035 доллара в час.

А при условии, что индивид будет не работать

$$\frac{\partial \mathbb{E}(wage_i^*|employment\_dummy_i = 0)}{\partial husband\_work\_hours} = \hat{\beta_h} + \hat{\rho} * \hat{\sigma} * (z_i'\hat{\gamma} * \tilde{\lambda_i} + \tilde{\lambda_i}^2) * \hat{\gamma_h} = -0.0003338386$$

То есть для этого индивида при увеличении количества рабочих часов у мужа на 100 возможная альтернативная зарплата при условии, что она не будет работать, упадет в среднем на 0.033 доллара в час.

Часть 4. Модель Ньюи.

Задание №4.1.*

Опишите преимущества и недостатки метода Ньюи по сравнению с методом Хекмана.

Метод Ньюи по сравнению с методом Хекмана

- Преимущества
  - Не используем предположение о совместном нормальном распределении ошибок, значит получаем более общее решение
  - Проще перевести на многомерный случай и условия участия
- Недостатки
  - Сложнее технически оценивать (нужно оценивать степени полиномов) + требует кросс-валидацию
  - Менее интерпретируем

Задание №4.2.***

Взяв за основу любую модель бинарного выбора (для простоты можно и параметрическую, например, логит), произвольную сглаживающую функцию и используя leave-one-out кросс-валидацию для подбора степени полинома, оцените модель Ньюи, описав осуществленные для ее построения шаги. Результат представьте в форме таблицы, содержащей оценки и бутстрапированные стандартные ошибки. Сравните оценки модели Ньюи и моделей Хекмана, основанных на двухшаговой процедуре и ММП.