

ПРАВИТЕЛЬСТВО РОССИЙСКОЙ ФЕДЕРАЦИИ

**Федеральное государственное автономное
образовательное учреждение высшего образования**

**Национальный исследовательский университет
«Высшая школа экономики»**

Факультет экономических наук
Образовательная программа «Экономика»

ДОМАШНЕЕ ЗАДАНИЕ 2

«Прикладная микроэконометрика»

Студент группы БЭК165
Зехов Матвей Сергеевич

Преподаватель:
Потанин Богдан Станиславович

Содержание

1	Теория и гипотезы	3
1.1	3
1.2	3
1.3	3
1.4	4
2	Обработка данных	4
3	Тобит-модель	4
3.1	4
3.2	4
3.3	4
3.4	5
3.5	5
3.6	5
3.7	6
3.8	6
4		7
4.1	7
4.2	7
4.3	7
4.4	7
5		8
5.1	8
5.2	8
5.3	9
5.4	9

1 Теория и гипотезы

1.1

☀ Выберите независимые переменные. Теоретически обоснуйте выбор каждой из них. Укажите предполагаемые направления эффектов. При этом вам понадобится как минимум одна непрерывная переменная и одна дамми-переменная (не рекомендуется брать больше трех различных переменных, не считая их нелинейных преобразований: квадрат, логарифм, перемножение с целью получения переменной взаимодействия и т.д.).

В качестве независимых переменных возьмём возраст, квадрат возраста, дамми-переменную, демонстрирующую наличие несовершеннолетних детей, а также переменную логарифм переменной $\ln y$. Последняя переменная отражает доход, который, по мнению индивида, должна иметь его семья, чтобы он считал, что живёт не хуже других.

Касательно возраста, очевидно, что предложение труда должно зависеть от возраста индивида, так как с возрастом у него могут изменяться условия. Например, молодёжь, которая ещё получает образование, имеет меньшее предложение труда, а опытные специалисты – большее. Ожидается, что эффект будет положительным. Эффект наличия несовершеннолетних детей также должен иметь положительный эффект, так как дети требуют, очевидно, больших расходов и индивиду придется больше работать. Желаемый доход отражает субъективный фактор мотивации. Предполагается, что чем больше денег нужно человеку для того, чтобы быть не хуже других, то тем больше он готов работать. Следовательно, эффект будет положительным. Предполагается, что этот показатель входит в уравнение нелинейно, а именно, логарифмически.

1.2

☀ Сформулируйте по крайней мере одну гипотезу о наличии эффекта взаимодействия и нелинейного эффекта (например, квадратичного). Теоретически обоснуйте выдвигаемые вами гипотезы. Включите нелинейные переменные и переменные взаимодействия в вашу модель.

Эффект взаимодействия наличия детей и мотивации отражает схожий с предыдущими эффект. Наличие детей будет повышать коэффициент перед мотивацией, так как для тех, кто имеет детей, также понадобится иметь больший доход, чтобы быть не хуже других. Эффект предполагается положительным. Квадратичный эффект возраста можно объяснить следующим образом. Предположим, что предложение труда зависит параболически от возраста. То есть, что есть точка максимума предложения труда в определённом возрасте, после которой предложение убывает для индивида. Следовательно, предполагается, что парабола будет ветвями вниз и эффект будет отрицательным.

1.3

☀ Определитесь с тем, будете ли вы изучать влияние на предложение труда среди мужчин и женщин в рамках единой модели, либо остановитесь лишь на одном из полов. Обоснуйте свой выбор теоретически.

Будем рассматривать выборку только совершеннолетних мужчин. Это вызвано тем, что наличие несовершеннолетних детей имеет разнонаправленный эффект на мужчин и женщин, поэтому рассматривать их в рамках одной модели выглядит бессмысленным. Несовершеннолетних индивидов не имеет смысла рассматривать, так как они вряд ли давали репрезентативные ответы о готовности работать.

1.4

☀ Определите границы усечения вашей зависимой переменной и теоретически обоснуйте их.

Так как в выборке нет ни одного совершеннолетнего мужчины, который бы ответил 0 часов (я сам удивился, но там ответ минимум 6 часов по 5 тысячам наблюдений), то граница усечения будет выше. Будем считать, что индивиды, предлагающие 30 и менее часов - это самозанятые, фрилансеры, безработные, работающие неофициально и тому подобные индивиды. Следовательно, усекать будем по 30 часам.

2 Обработка данных

3 Тобит-модель

3.1

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

$$L = \prod_{y_t=0} \left(1 - \Phi \left(\frac{\mathbf{x}'_t \boldsymbol{\beta}}{\sigma} \right) \right) \prod_{y_t > 0} \frac{1}{\sqrt{2\pi}\sigma} \exp \left(-\frac{1}{2\sigma^2} (y_t - \mathbf{x}'_t \boldsymbol{\beta})^2 \right)$$

	Estimate	Std. Error	z value	Pr(> z)
(Intercept)	15.3498	9.5342	1.6100	0.1074
age	0.4865	0.2195	2.2165	0.0267
I(age^2)	-0.0057	0.0024	-2.3929	0.0167
child	23.4039	10.4644	2.2365	0.0253
lmotiv	1.7636	0.6987	2.5243	0.0116
I(lmotiv * child)	-1.8355	0.9058	-2.0264	0.0427
Log(scale)	2.5198	0.0193	130.2517	0.0000

Таблица 1: Результаты оценки tobit-модели

3.2

Как можно видеть из Таблицы 1, все переменные значимы на 5%-ом уровне, а также за исключением переменной взаимодействия имеют предполагаемые знаки. Не могу дать содержательной интерпретации знаку этого коэффициента.

3.3

☀ Запишите формулы, по которым можно рассчитать предельные эффекты в тобит модели для переменной, входящей линейно, в отношении:

- А) $E(y^*)$
- Б) $E(y)$
- В) Вероятности того, что индивид работает

1.

$$\mathbb{E}(Y_i^*) = x_i' \beta$$

$$\frac{d\mathbb{E}(Y_i^*)}{dx^j} = \beta^j$$

2.

$$\frac{d\mathbb{E}(Y_i)}{dx^j} = \Phi\left(\frac{x_i' \beta - 30}{\sigma}\right) \beta^j$$

3.

$$\mathbb{P}\{y^* > 30\} = \Phi\left(\frac{x_i' \beta - 30}{\sigma}\right)$$

$$\frac{d\mathbb{P}\{y^* > 30\}}{dx^j} = \phi\left(\frac{x_i' \beta - 30}{\sigma}\right) \beta^j$$

3.4

☀️ Проинтерпретируйте предельный эффекты для всех независимых переменных входящих нелинейно или имеющих взаимодействие на $\mathbb{E}(Y_i^*)$

Так как выписывать формулы никто и не просит, то я и не буду. Расчёты будут для медианного индивида. Его характеристики представлены в Таблице 2, а предельные эффекты - в Таблице 3.

Возраст	Наличие несов. детей	Мотивация
44.00	1.00	11.51

Таблица 2: Характеристики медианного индивида

Возраст	Наличие несов. детей	Мотивация
-0.01	-18.86	-1.91

Таблица 3: Предельные эффекты медианного индивида

При увеличении возраста медианного индивида на 1 год предложение труда снижается на 0.01 часа в неделю. При наличии несовершеннолетних детей предложение труда снижается на 18.86 часов. При увеличении "мотивации" на 1% предложение труда снижается на 1.91 часа в неделю.

3.5

☀️ Проинтерпретируйте предельный эффекты для всех независимых переменных входящих линейно на $\mathbb{E}(y)$.

Линейно входящих переменных нет.

3.6

☀️ Проинтерпретируйте предельный эффекты для всех независимых переменных на вероятность того, что индивид работает.

Предельные эффекты представлены в Таблице 4. При увеличении возраста на 1 год вероятность занятости уменьшается на 0.00001. При наличии несовершеннолетних детей вероятность уменьшается на 0.016. При увеличении "мотивации" на 1% вероятность снижается на 0.00167.

Возраст	Наличие несов. детей	Мотивация
-0.00001	-0.01647	-0.00167

Таблица 4: Предельные эффекты вероятности занятости медианного индивида

3.7

☀Предскажите значения $E(y^*)$, $E(y)$ и вероятности занятости для индивида с вашими характеристиками, выписав формулу, по которой осуществлялся расчет.

Мои характеристики представлены в Таблице 5. Прогнозные значения представлены в Таблице 6.

Формулы:

v - вектор моих характеристик

1. $\mathbb{E}(\hat{Y}^*) = v' \hat{\beta}$
2. $\mathbb{E}(\hat{Y}) = \Phi\left(\frac{v' \hat{\beta} - 30}{\hat{\sigma}}\right) v' \hat{\beta} + \hat{\sigma} \phi\left(\frac{v' \hat{\beta} - 30}{\hat{\sigma}}\right)$
3. $\mathbb{P}\{\hat{Y}^* > 30\} = \Phi\left(\frac{v' \hat{\beta} - 30}{\hat{\sigma}}\right)$

	Intercept	Возраст	Возраст**2	Несов. дети	Логарифм з.п.	Дети*Логарифм з.п.
1	1.00	21.00	441.00	0.00	12.21	0.00

Таблица 5: Мои характеристики

$\mathbb{E}(\hat{Y}^*)$	$\mathbb{E}(\hat{Y})$	$\mathbb{P}\{\hat{Y}^* > 30\}$
44.58	39.42	0.88

Таблица 6: Прогнозы

3.8

☀Выведите, описав процесс вывода в тексте работы, формулу для расчета предельного эффекта переменной, входящей нелинейно, на $\mathbb{E}(Y)$ Вспомним формулу:

$$\mathbb{E}(\hat{Y}) = \Phi\left(\frac{x' \hat{\beta} - 30}{\hat{\sigma}}\right) x' \hat{\beta} + \hat{\sigma} \phi\left(\frac{x' \hat{\beta} - 30}{\hat{\sigma}}\right)$$

При наших переменных возьмём производную по возрасту, воспользовавшись тем, что $\phi' = -x\phi$

$$\begin{aligned} \frac{d\mathbb{E}(\hat{Y}_i)}{dage} &= \left(\frac{\beta_1 + 2\beta_2 age_i}{\hat{\sigma}}\right) \phi\left(\frac{x'_i \hat{\beta} - 30}{\hat{\sigma}}\right) x'_i \hat{\beta} + \Phi\left(\frac{x'_i \hat{\beta} - 30}{\hat{\sigma}}\right) (\beta_1 + 2\beta_2 age_i) + \\ &\quad \hat{\sigma} \left(\frac{30 - x'_i \hat{\beta}}{\hat{\sigma}}\right) \phi\left(\frac{x'_i \hat{\beta} - 30}{\hat{\sigma}}\right) = \\ &\quad \phi(\gamma_i) \left(\left(\frac{\beta_1 + 2\beta_2 age_i}{\hat{\sigma}}\right) x'_i \hat{\beta} + \hat{\sigma} (-\gamma_i) + \frac{\beta_1 + 2\beta_2 age_i}{\lambda_i}\right) \end{aligned}$$

где $\lambda_i = \frac{\phi(\gamma_i)}{\Phi(\gamma_i)}$, $\gamma_i = \frac{x'_i \hat{\beta} - 30}{\hat{\sigma}}$

4

4.1

☀ Проверьте, можно ли исключить из модели эффект взаимодействия, предварительно записав нулевую гипотезу и статистику теста, а также указав ее распределение.

$$H_0 : \beta_{lmotiv*child} = 0$$

$$H_A : \beta_{lmotiv*child} \neq 0$$

Оценки ММП асимптотически нормальны, так что гипотеза проверется Z-тестом. Статистика:

$$\frac{\hat{\beta}_{lmotiv*child} - 0}{\hat{\sigma}_{\beta}} \sim N(0,1)$$

p-value было представлено в Таблице 1. Основная гипотеза отвергается на 5%-ом уровне.

4.2

☀ Проверьте, можно ли оценивать совместную модель для тех, кто состоит в браке и тех, кто нет, предварительно записав нулевую гипотезу и статистику теста, а также указав ее распределение.

Проверим аналогично LR-тестом.

$$H_0 :$$

$$H_A : , .$$

Оценим три модели и получим статистику теста:

$$2(l_{UR1} + l_{UR2} - l_R) \sim \chi^2_7$$

Семь степеней свободы, так как каждая модель имеет по 6 коэффициентов и плюс стандартное отклонение. Нулевая гипотеза отвергается на любом разумном уровне значимости.

4.3

☀ Опишите, к чему могут приводить, во-первых, гетероскедастичность, а во-вторых, нарушение допущения о распределении случайной ошибки в тобит модели

Гетероскедастичность может привести как к неэффективности оценок, так и к несостоятельности, так как если правдоподобие будет построено исходя из гомоскедастичности, то будет максимизировано не то правдоподобие. Оценки будут смещёнными и несостоятельными.

При нарушении предпосылки о распределении случайной ошибки опять же, будет максимизировано неверное правдоподобие и, как следствие, смещённые и несостоятельные оценки.

4.4

Запишите функцию правдоподобия для тобит модели с гетероскедастичной случайной ошибкой и формально опишите тестирование гипотезы о наличии гетероскедастичности. Для тех, кто работает в R, дополнительно следует оценить параметры данной модели

используя функцию с семинара и представить результаты в форме таблицы, а также сделать вывод о наличии или отсутствии гетероскедастичности.

$$L = \prod_{y_i=0} \left(1 - \Phi \left(\frac{\mathbf{x}'_i \boldsymbol{\beta}}{\sigma_i} \right) \right) \prod_{y_i>30} \frac{1}{\sqrt{2\pi}\sigma_i} \exp \left(-\frac{1}{2\sigma_i^2} (y_i - \mathbf{x}'_i \boldsymbol{\beta})^2 \right)$$

Для проверки гипотезы оценим модель с учётом гетероскедастичности и без, и сравним их LR-тестом. Ограничения будут наложены на уравнение, описывающее дисперсию. Количество ограничений будет равно количеству переменных, от которых зависит дисперсия модели. Обозначим эту величину за q .

$$H_0 : \sigma_i = \sigma$$

$$H_A : \exists \text{ i s.t. } \sigma_i \neq \sigma$$

$$2(l_{UR} - l_R) \sim \chi_q^2$$

Предположим, что гетероскедастичность порождается возрастом. Например, у молодёжи может сильнее колебаться предложение труда в связи с учёбой, личной жизнью, хобби, социальной активностью и прочим. Хотя откуда мне знать про всё, кроме первого, я же в вышке учусь. Гипотеза отвергается на любом разумном уровне значимости. Результаты оценки модели с учётом гетероскедастичности представлены в Таблице 7. Как хорошо видно, коэффициент при возрасте в уравнении дисперсии незначим.

	Estimate	Std. Error	z value	Pr(> z)
(Intercept)	15.2810	9.4520	1.6167	0.1059
age	0.4968	0.2195	2.2635	0.0236
I(age^2)	-0.0058	0.0024	-2.4503	0.0143
child	23.0467	10.4263	2.2104	0.0271
lmotiv	1.7458	0.6890	2.5336	0.0113
I(lmotiv * child)	-1.8017	0.9026	-1.9962	0.0459
(scale)_(Intercept)	2.5804	0.0844	30.5908	0.0000
(scale)_age	-0.0014	0.0019	-0.7406	0.4589

Таблица 7: Результаты оценки модели с гетероскедастичностью

5

5.1

☀Оцените две дополнительные модели: с помощью МНК и усеченной регрессии, представив результаты оценивания в форме таблицы, а также выписав функцию правдоподобия для усеченной регрессии.

Функция правдоподобия усечённой модели:

$$l = -\frac{n}{2} (\ln(2\pi) + \ln \sigma^2) - \frac{1}{2\sigma^2} \sum_i (y_i - \mathbf{x}'_i \boldsymbol{\beta})^2 - \sum_i \ln \left[1 - \Phi \left(\frac{30 - \mathbf{x}'_i \boldsymbol{\beta}}{\sigma} \right) \right]$$

Результаты оценки усечённой модели представлены в Таблице 8. Результаты оценки линейной модели представлены в Таблице 9.

	Estimate	Std. Error	z value	Pr(> z)
(Intercept)	2.4252	23.5301	0.1031	0.9179
age	0.9390	0.5778	1.6251	0.1041
I(age^2)	-0.0113	0.0063	-1.7901	0.0734
child	13.2390	24.4356	0.5418	0.5880
lmotiv	1.3219	1.6769	0.7883	0.4305
I(lmotiv * child)	-0.8876	2.1186	-0.4190	0.6752
sigma	17.6455	0.8673	20.3444	0.0000

Таблица 8: Результаты оценки усечённой модели

	Estimate	Std. Error	t value	Pr(> t)
(Intercept)	15.7208	9.6692	1.6259	0.1042
age	0.4490	0.2221	2.0222	0.0433
I(age^2)	-0.0053	0.0024	-2.1951	0.0283
child	23.4596	10.6378	2.2053	0.0276
lmotiv	1.7817	0.7090	2.5128	0.0121
I(lmotiv * child)	-1.8257	0.9208	-1.9827	0.0476

Таблица 9: Результаты оценки линейной модели

5.2

☀ Какие оценки являются более эффективными: тобит модели или у усеченной регрессии. Ответ обоснуйте.

Здесь следует сказать, что в Тобит-модели всё же содержится некоторая информация о усечённых наблюдениях, хотя и искажённая. В усечённой модели же эти данные отсутствуют вовсе, и корректировка происходит исходя из предположений о распределении данных. Следовательно, первые оценки будут более эффективными, так как они учитывают больше информации. Это видно из результатов оценки. В усечённой регрессии большая часть переменных стала незначимой.

5.3

☀ Выберите лучшую модель на основании критерия AIC.

По AIC нет смысла сравнивать, так как модели оценены на разном количестве наблюдений.

5.4

☀ Выберите лучшую модель по критерию MSE, посчитанному на тестовой выборке

Отведём на тестовую часть 0.2 выборки и переоценим модели на оставшихся данных. По результатам вневыборочного прогноза из Таблицы 10 можно сказать, что усечённая модель даёт слишком смещённые прогнозы. Линейная и Tobit-модели почти эквивалентны, но линейная слегка лучше. Её и признаем оптимальной.

Усечённая	Линейная	Tobit
241.1	153.9	154.2

Таблица 10: MSE вневыборочного прогноза