Национальный исследовательский университет
“Высшая школа экономики”

Факультет экономических наук

Веселова Арина Олеговна

Домашнее задание № 1

Отчет студентки 4 курса бакалавриата группы БЭК201

по Моделям с качественными и ограниченными зависимыми переменными

Москва 2023

Часть 1. Теория и гипотезы.

Задание № 1.1. Выберите независимые переменные. Кратко теоретически обоснуйте выбор каждой из них: не обязательно со ссылками на литературу, достаточно здравого смысла. Укажите и кратко обоснуйте предполагаемые направления эффектов. При этом вам понадобится как минимум одна непрерывная переменная (например, возраст или доход) и одна дамми переменная (например, половая принадлежность или брак). Не рекомендуется брать больше трех различных независимых переменных, не считая их нелинейных преобразований: квадрат, логарифм, перемножение с целью получения переменной взаимодействия и т.д.

Независимые переменные:

1. age – именно молодые люди и люди среднего возраста склонны к оформлению подписки в онлайн-кинотеатрах ввиду их большей технической оснащенности и осведомленности, гораздо меньшая же доля пожилых людей пользуется телефоном, телевизором с веб-системами и Интернетом в целом, что понижает вероятность оформления ими подписки.
2. series –люди, которые просмотрели большое количество сериалов за год, вероятнее оформят подписку в онлайн-кинотеатре, поскольку это один из основных каналов просмотра сериалов и факт активного увлечения сериалами будет косвенно свидетельствовать об их нужде в подписке.
3. TV – если тут речь о телевидении: активный просмотр телепрограмм может косвенно свидетельствовать о возрасте человека, поскольку молодое поколение все реже таким способом проводит свободное время, тогда если человек редко смотрит телевизор, он должен каким-то иным способом проводить свое свободное время и таким способом может стать Интернет, социальные сети, просмотр видео, сериалов, фильмов, последние из которых чаще всего смотрят с помощью онлайн-кинотеатров. Таким образом, чем меньше времени человек проводит за телевизором, тем выше шанс, что его времяпрепровождением может стать онлайн-кинотеатр и, как следствие, оформление подписки на него

Задание 1.2. Сформулируйте по крайней мере одну гипотезу о наличии эффекта взаимодействия и **еще** одну о наличии нелинейного эффекта (например, квадратичного). Теоретически обоснуйте выдвигаемые вами гипотезы. Включите соответствующие переменные в вашу модель. При этом переменная, входящая нелинейно, должна иметь и линейную часть, например, $\beta * X + \beta * X^2$.

Гипотезы:

1. Гипотеза о наличии эффекта взаимодействия: age*series. Люди, любящие смотреть сериалы, могут по-разному себя вести в зависимости от возраста. Более взрослые люди, которые любят смотреть сериалы, вероятнее будут пользоваться легальными способами такого времяпрепровождения, то есть чаще оформлять подписку на онлайн-кинотеатр, молодые же, имея меньшие финансовые возможности и большую «прозорливость», будут пользоваться пиратскими сайтами и не оформлять подписку.
2. age^2 – зависимость между вероятностью оформления подписки и возрастом может быть гиперболического вида с ветвями вниз, поскольку до определенного возраста вероятность будет расти (это можно объяснить финансовыми возможностями и изменениями интересов), а затем падать, поскольку с какого-то возраста техническая оснащенность будет падать, а с ней и нужда в онлайн-кинотеатрах.

Часть 2. Линейно-вероятностная модель.

Задание 2.1. Оцените линейно-вероятностную модель, предварительно записав регрессионное уравнение. Укажите оцениваемые параметры и метод получения оценок. Результат представьте в форме таблицы (можно, например, использовать выдачу из stata, R или python).

$$P(sub = 1) = \beta_0 + \beta_1 age + \beta_2 series + \beta_3 TV + \beta_4 age^2 + \beta_5 age \cdot series + \varepsilon$$

Таким образом, оцениваются коэффициенты при объясняющих переменных $\beta_0, \beta_1, \beta_2, \beta_3, \beta_4, \beta_5$ с помощью МНК.

```
Coefficients:
                  Estimate Std. Error t value Pr(>|t|)
(Intercept)      1.376e-01  5.539e-02   2.484 0.013023 *
age              5.841e-03  1.721e-03   3.395 0.000692 ***
series           4.011e-02  6.255e-03   6.412 1.56e-10 ***
TV              -2.813e-01  1.443e-02 -19.494  < 2e-16 ***
I(age^2)        -2.157e-05  1.352e-05  -1.595 0.110686
I(age * series) -2.074e-04  9.654e-05  -2.148 0.031722 *
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1
```

Табл. 1. Результаты оценивания линейно-вероятностной модели.

Задание 2.2. Перечислите основные недостатки линейно-вероятностной модели. Напишите, можно ли интерпретировать оценки коэффициентов, их значимость (с использованием обычной оценки ковариационной матрицы), коэффициент детерминации и F-статистику? Если да, то приведите интерпретацию, а если нет, то объясните (без непосредственной реализации), почему она в данном случае невозможна и предложите альтернативный способ оценки качества модели.

Основные недостатки:

- оцененные значения вероятности могут оказаться больше 1 или меньше 0,
- распределение случайного члена не является нормальным
- гетероскедастичность

Оценки коэффициентов из такой модели можно интерпретировать только для переменной TV (то есть с линейной зависимостью):

- если индивид смотрит телевизор не реже раза в неделю, то вероятность оформления подписки в онлайн-кинотеатре уменьшается на 0.2813

Остальные же коэффициенты нужно интерпретировать с помощью предельных эффектов.

Из-за гетероскедастичности значимость, коэффициент детерминации и F-статистику нельзя интерпретировать. Чтобы избавиться от проблемы, можно было бы использовать взвешенный МНК или использовать оценки стандартных ошибок в форме Уайта.

Задание 2.3. Оцените и проинтерпретируйте, независимо от значимости, предельные эффекты на вероятность подписки каждой из используемых вами независимых переменных, предварительно записав формулы, по которым осуществлялся расчет. Результат представьте в форме таблицы, где для переменных, входящих нелинейно, рассчитан средний предельный эффект. Также, для этих переменных должно быть указано, при каких значениях независимой переменной их предельный эффект является положительным, а при каких — отрицательным.

Для переменной TV:

$$\frac{\partial P(\widehat{sub} = 1)}{\partial TV} = P(\widehat{sub} = 1|TV = 1) - P(\widehat{sub} = 1|TV = 0) = \widehat{\beta_3} = -0.2813$$

Для переменной age:

$$\frac{\partial P(\widehat{sub} = 1)}{\partial age} = \widehat{\beta_1} + 2\widehat{\beta_4}age + \widehat{\beta_5}series$$
$$= 0.005841 - 0.00004313age - 0.0002074series$$

Для переменной series:

$$\frac{\partial P(\widehat{sub} = 1)}{\partial series} = \widehat{\beta_2} + \widehat{\beta_5}age = 0.04011 - 0.0002074age$$

Переменная	Предельный эффект	Значения переменной, при которых пр. эффект положительный	Значения переменной, при которых пр. эффект отрицательный
TV	-0.2813	-	∀
age	0.0022	4313age+20740series<584100	4313age+20740series>584100
series	0.0276	Age<193.39	Age>193.39

Табл. 2. Предельные эффекты на вероятность подписки каждой из независимых переменных.

Задание 2.4*. Протестируйте гипотезы о значимости коэффициентов с помощью бутстрапа. Результат представьте в форме таблицы. При этом предварительно (словами или самостоятельно нарисованной схемой) опишите алгоритм, который вы использовали для построения бутстрапированных доверительных интервалов.

Алгоритм бутстрапа: для получения бутстрапированных оценок мы будем на протяжении n итераций (сами подбираем гиперпараметр, в нашем случае, возьмем 100) создавать новую выборку из строк изначального датасета его же размерностью (причем вероятности взять i-ую и j-ую строки исходного датасета равны), тем самым получим для каждого параметра 100 оценок, каждую из которых усредним, и выведим бустрапированные оценки.

	Feature	Left_quantile	Right_quantile
1	age	3.073784e-03	8.792478e-03
2	series	2.776300e-02	5.187486e-02
3	TV	-3.255191e-01	-2.553980e-01
4	age^2	-4.938171e-05	5.406368e-06
5	age * series	-4.165429e-04	-2.142056e-05

Табл. 3. 95% бутстрапированные доверительные интервалы для коэффициентов линейно-вероятностной модели.

Таким образом, незначимым на уровне значимости 5% оказался параметр age^2, поскольку по этому параметру в бутстрапированный ДИ входит 0, остальные же признаки оказались значимыми.

Задание 2.5.** Протестируйте гипотезы о значимости коэффициентов используя состоятельную (скорректированную на гетероскедастичность) оценку асимптотической ковариационной матрицы. Результат представьте в форме таблицы (в том числе с p-value), предварительном выписав используемую для расчетов формулу.

Будем использовать робастные ошибки в форме Уайта:

$$Var\left(\hat{\beta}\middle|X\right) = (X'X)^{-1}X'\hat{\Sigma}X(X'X)^{-1}, \qquad \text{где } \hat{\Sigma} = diag(\hat{\epsilon}_1^2, \hat{\epsilon}_2^2, \dots, \hat{\epsilon}_n^2), \hat{\epsilon}_i^2 = y - \hat{y}$$

Coefficients:

	Estimate	Std. Error	t value	Pr(>|t|)	CI Lower	CI Upper	DF
(Intercept)	1.376e-01	5.290e-02	2.601	9.325e-03	3.388e-02	2.413e-01	4994
age	5.841e-03	1.689e-03	3.458	5.493e-04	2.529e-03	9.153e-03	4994
series	4.011e-02	6.092e-03	6.584	5.061e-11	2.817e-02	5.205e-02	4994
TV	-2.813e-01	1.430e-02	-19.674	4.689e-83	-3.093e-01	-2.533e-01	4994
I(age^2)	-2.157e-05	1.351e-05	-1.596	1.105e-01	-4.807e-05	4.920e-06	4994
I(age * series)	-2.074e-04	9.539e-05	-2.174	2.972e-02	-3.944e-04	-2.042e-05	4994

Multiple R-squared: 0.08195 , Adjusted R-squared: 0.08103
F-statistic: 100 on 5 and 4994 DF, p-value: < 2.2e-16

Табл. 4. Линейно-вероятностная модель с робастными ошибками в форме Уайта.

Так, на 5%-ом уровне значимости значимыми оказались параметры age, series, TV, age*series и константы; age^2 оказался же незначимым, поскольку p-value = 0.11 > 0.05.

Часть 3. Пробит модель.

Задание 3.1. Оцените пробит модель, предварительно записав максимизируемую функцию правдоподобия (поясните все используемые обозначения), указав оцениваемые параметры и метод получения оценок, а также их основные свойства. Результат представьте в форме таблицы (можно, например, использовать выдачу из stata, R или python).

$y_i = \begin{cases} 1, & \text{если } y_i^* > 0 \\ 0, & \text{если } y_i^* \le 0 \end{cases}$, где $y_i^* = x_i'\beta + \epsilon_i$ и $\epsilon_i \sim N(0,1)$

Тогда $P(y_i = 1) = \Phi(x_i'\beta) = \int_{-\infty}^{x_i'\beta} \frac{1}{\sqrt{2\pi}} e^{-\frac{t^2}{2}} dt$

С помощью ММП будут оцениваться коэффициенты модели β:

$$l = \ln(L) = \ln\left(\prod_{i=1}^{N} [\Phi(x_i'\beta)]^{y_i} [1 - \Phi(x_i'\beta)]^{1-y_i}\right)$$
$$= \sum_{i=1}^{N} \left[y_i \ln\left(\Phi(x_i'\beta)\right) + (1 - y_i) \ln\left(1 - \Phi(x_i'\beta)\right)\right] \to \max_{\beta}$$

Свойства:

- состоятельность
- асимптотическая несмещенность
- асимптотическая эффективность
- асимптотическая нормальность
- инвариантность

```
Coefficients:
                  Estimate Std. Error z value Pr(>|z|)
(Intercept)     -1.073e+00  1.629e-01  -6.585 4.54e-11 ***
age              1.738e-02  4.986e-03   3.486  0.00049 ***
series           1.188e-01  1.824e-02   6.513 7.36e-11 ***
TV              -7.823e-01  4.181e-02 -18.709  < 2e-16 ***
I(age^2)        -6.371e-05  3.883e-05  -1.641  0.10085
I(age * series) -6.380e-04  2.779e-04  -2.296  0.02170 *
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

(Dispersion parameter for binomial family taken to be 1)

    Null deviance: 6538.8  on 4999  degrees of freedom
Residual deviance: 6116.6  on 4994  degrees of freedom
AIC: 6128.6
```

Табл. 5. Оценки коэффициентов пробит-модели.

Задание 3.2. Проинтерпретируйте оценки коэффициентов для каждой независимой переменной. Поясните, как полученные результаты соотносятся с высказанными вами ранее предположениями.

Поскольку при age^2 p-value = 0.101, то на 5%-ом уровне значимости параметр не значим, то есть между вероятностью подписки и возрастом нет квадратичной зависимости.

Остальные коэффициенты значимы на выбранном уровне значимости:

- с увеличением возраста или количества просмотренных сериалов за год вероятность оформления подписки повышается, как мной и предполагалось, однако с одновременным возрастанием и возраста, и количества сериалов вероятность подписки понижается, вероятно, поскольку здесь будут появляться разные ТВ-сериалы, которые активнее смотрят пожилые люди, не интересующиеся онлайн-кинотеатрами
- как и ожидалось, активный просмотр ТВ показывает слабый интерес к онлайн-кинотеатрам и снижает вероятность оформления подписки

Задание 3.3. Оцените вероятность наличия подписки для индивида с произвольными (например, вашими) характеристиками. Запишите формулу, по которой осуществлялся расчет (подставьте в нее полученные реализации оценок).

Оценивать вероятность наличия подписки будем для моих характеристик:

Параметр	age	series	TV	age^2	age*series
Значение	21	10	0	441	210

Тогда вероятность:

$$\begin{aligned}\hat{P}(sub = 1) &= \Phi(-1.073 + 0.017age + \ 0.119series - 0.782TV + 0.000064age^2 \\ &\quad - 0.00064age \cdot series) \\ &= \Phi(-1.073 + 0.017 * 21 + \ 0.119 * 10 - 0.782 * 0 + 0.000064 * 441^2 \\ &\quad - 0.00064 * 210) = 0.625\end{aligned}$$

Задание 3.4. Для произвольных непрерывной и бинарной независимых переменных оцените средний предельный эффект на вероятность наличия подписки, предварительно записав формулы (с подставленными реализациями оценок), по которым осуществлялся расчет. Результат представьте в форме таблицы.

Для переменной age:

$$\frac{\partial P(\widehat{sub} = 1)}{\partial age} = f(x'\hat{\beta})(\widehat{\beta_1} + 2\widehat{\beta_4}age + \widehat{\beta_5}series)$$

$$= \frac{1}{\sqrt{2\pi}} e^{-\frac{\left(-1.073+0.017age+\ 0.119series-0.782TV+0.000064age^2-0.00064age\cdot series\right)^2}{2}}(0.017 + \ 2 * 0.000064age \ - 0.00064series)$$

Для переменной TV:

$$\frac{\partial P(\widehat{sub}=1)}{\partial TV} = P(\widehat{sub}=1|TV=1) - P(\widehat{sub}=1|TV=0)$$
$$= \Phi(-1.073 + 0.017age + 0.119series - 0.782 + 0.000064age^2 - 0.00064age \cdot series)$$
$$- \Phi(-1.073 + 0.017age + 0.119series + 0.000064age^2 - 0.00064age \cdot series)$$

Переменная	Предельный эффект
TV	-0.273
age	0.0022

Табл. 6. Предельные эффекты на вероятность подписки каждой из независимых переменных.

Задание 3.5. Посчитайте долю верных предсказаний и сопоставьте её с результатом наивного прогноза и линейно-вероятностной модели. Сделайте вывод о предсказательной силе пробит модели.

```
probit  67.30
linprob 67.26
naive   63.92
```

Табл. 7. Доля верных предсказаний для пробит, линейно-вероятностной модели и наивного прогноза.

Таким образом, пробит-модель показала самую высокую предсказательную силу.

Задание 3.6. На уровне значимости 5% проверьте гипотезу о том, что предельный эффект на вероятность наличия подписки по произвольной (на ваш выбор) независимой переменной является значимым для индивида с произвольными характеристиками.

$$H_0: \frac{\partial P(\widehat{sub}=1|X_i)}{\partial TV} = 0$$

Выведем информацию о характеристиках оцененных ранее предельных эффектах:

```
factor     AME     SE       z      p   lower   upper
   age -0.0015 0.0050 -0.3016 0.7630 -0.0113  0.0083
series -0.0890 0.0410 -2.1679 0.0302 -0.1694 -0.0085
    TV  0.0694 0.2849  0.2435 0.8076 -0.4890  0.6278
```

Табл. 8. Информация об оценках предельных эффектов пробит-модели.

Таким образом, видим, что p-value для TV равен 0.81, что больше любого разумного уровня значимости, значит, нулевая гипотеза не отвергается и предельный эффект на вероятность наличия подписки по независимой переменной TV является незначимым.

Проверим гипотезу с помощью бутстрапированного доверительного интервала:

```
        2.5%       97.5%
-0.05353349  0.05278783
```

Поскольку 0 входит в ДИ, то предельный эффект на вероятность наличия подписки по независимой переменной age является незначимым.

Задание 3.7*. Повторите предыдущий пункт для переменной, имеющей взаимодействие.

$$H_0: \frac{\partial P(\widehat{sub} = 1|X_i)}{\partial age} = 0$$

Предельный эффект для переменной age:

$$\frac{\partial P(\widehat{sub} = 1)}{\partial age} = f(x'\hat{\beta})(\widehat{\beta_1} + 2\widehat{\beta_4}age + \widehat{\beta_5}series)$$

Оценим значимость с помощью бутстрапированного доверительного интервала с моими характеристиками:

Параметр	age	series	TV	age^2	age*series
Значение	21	10	0	441	210

```
        2.5%         97.5%
0.000000e+00 4.671711e-67
```

Поскольку 0 не входит в доверительный интервал, то нулевая гипотеза отвергается и предельный эффект на вероятность наличия подписки по независимой переменной age является значимым.

Часть 4. Тестирование корректности спецификации пробит модели.

Задание 4.1. При помощи LM-теста проверьте гипотезу о соблюдении допущения о нормальном распределении случайных ошибок в пробит модели. Укажите, к каким негативным последствиям может привести нарушение данного допущения.

Предположим, что в пробит модели случайные ошибки имеют распределение Пирсона:

$$P(y = 1) = \Phi(x'\beta + \theta_1(x'\beta)^2 + \theta_2(x'\beta)^3)$$

$H_0: \theta_1 = \theta_2 = 0$, то есть ошибки распределены нормально (ограниченная модель)

Тестовая статистика:

$$LM = \left[\frac{\partial l_{UR}}{\partial \beta}\right]' I^{-1}(\widehat{\beta_R}) \left[\frac{\partial l_{UR}}{\partial \beta}\right] \sim \chi^2\ (2)$$

Если же нулевая гипотеза будет отвергнута, то это значит, что спецификация модели подобрана неверно и ее результаты не подлежат интерпретации.

После реализации LM-теста получаем p-value = 0.5339338, так как оно больше любого разумного уровня значимости, то нулевая гипотеза не отвергается и случайные ошибки распределены нормально.

Задание 4.2. Предположите, какие переменные могут влиять на дисперсию случайной ошибки. При этом по крайней мере одна переменная должна входить и в линейный индекс основного уравнения, и в линейный индекс уравнения дисперсии. При помощи LR теста проверьте гипотезу о гомоскедастичности случайных ошибок. Запишите, к каким негативным последствиям может привести нарушение данного допущения. Объясните преимущество LM теста над LR тестом в данном случае.

Предположим, что возраст и частый просмотр телевизора оказывают влияние на дисперсию случайной ошибки, то есть $\varepsilon \sim N(0, g(\gamma age + \mu TV)^2)$.

$$H_0: \gamma = \mu = 0$$

Тестовая статистика:

$$LR = -2(l_R(\widehat{\beta_R}) - l_{UR}(\widehat{\beta_{UR}})) \sim \chi^2(2)$$

После реализации LR-теста получаем p-value = 0.7211, так как оно больше любого разумного уровня значимости, то нулевая гипотеза не отвергается и случайные ошибки гомоскедастичны.

Задание 4.3. Для модели с гетероскедастичной случайной ошибкой рассчитайте предельный эффект на вероятность подписки и на дисперсию случайной ошибки по переменной, входящей и в основное уравнение, и уравнение дисперсии. Предварительно запишите формулы, по которым осуществляется расчет.

Оценивать вероятность наличия подписки будем для моих характеристик:

Параметр	age	series	TV	age^2	age*series
Значение	21	10	0	441	210

Для переменной age:

Для оценки вероятности нужно привести линейный индекс индивида к нормальному стандартному распределению (где $\sigma = e^{\gamma age + \mu TV}$):

$$\frac{x'\hat{\beta}}{\sigma} = \frac{\widehat{\beta_0} + \widehat{\beta_1} age + \widehat{\beta_2} series + \widehat{\beta_3} TV + \hat{\beta}_4 age^2 + \widehat{\beta_5} age \cdot series}{\sigma}$$
$$= \frac{\widehat{\beta_0} + \widehat{\beta_1} age + \widehat{\beta_2} series + \widehat{\beta_3} TV + \hat{\beta}_4 age^2 + \widehat{\beta_5} age \cdot series}{e^{\widehat{\gamma} age + \widehat{\mu} TV}}$$

Тогда предельный эффект на вероятность подписки:

$$\frac{\partial P(\widehat{sub} = 1)}{\partial age} = f\left(\frac{x'\hat{\beta}}{\sigma}\right) \cdot$$

$$\cdot \frac{\left(\widehat{\beta_1} + 2\widehat{\beta_4} age + \widehat{\beta_5} series\right)\sigma - (\widehat{\beta_0} + \widehat{\beta_1} age + \widehat{\beta_2} series + \widehat{\beta_3} TV + \hat{\beta}_4 age^2 + \widehat{\beta_5} age \cdot series)\gamma\sigma}{\sigma^2}$$

$$= f\left(\frac{x'\hat{\beta}}{\sigma}\right) \cdot$$

$$\cdot \frac{\left(\widehat{\beta_1} + 2\widehat{\beta_4} age + \widehat{\beta_5} series\right) - \left(\widehat{\beta_0} + \widehat{\beta_1} age + \widehat{\beta_2} series + \widehat{\beta_3} TV + \hat{\beta}_4 age^2 + \widehat{\beta_5} age \cdot series\right)\gamma}{\sigma}$$

Таким образом, предельный эффект на вероятность подписки равен 0.0034.

Предельный эффект на дисперсию случайной ошибки:

$$\frac{\partial e^{2(\gamma age + \mu TV)}}{\partial age} = 2\gamma e^{2(\gamma age + \mu TV)}$$

Таким образом, предельный эффект на дисперсию случайной ошибки равен -0.0021.

Задание 4.4. Для переменной, входящей в линейный индекс нелинейно, при помощи LR теста проверьте гипотезы о том, что:

1) Коэффициент при линейной части равняется нулю

Проверять будем для переменной series:

$$H_0: \beta_2 = 0$$

Тестовая статистика:

$$LR = -2(l_R(\widehat{\beta_R}) - l_{UR}(\widehat{\beta_{UR}}))\sim\chi^2(1)$$

После реализации LR-теста получаем p-value = 1, так как оно больше любого разумного уровня значимости, то нулевая гипотеза не отвергается и коэффициент незначим.

2) Оба коэффициента равняются нулю

$$H_0: \begin{cases}\beta_2 = 0\\ \beta_5 = 0\end{cases}$$

Тестовая статистика:

$$LR = -2(l_R(\widehat{\beta_R}) - l_{UR}(\widehat{\beta_{UR}}))\sim\chi^2(2)$$

После реализации LR-теста получаем, что p-value стремится к 0, так как оно меньше любого разумного уровня значимости, то нулевая гипотеза отвергается и коэффициент значим.

3) Коэффициент при линейной части совпадает по знаку и в k раз больше (по модулю), чем при нелинейной, где $k \neq 0$ можно выбрать произвольным, указав выбранное значение.

Пусть k = 3:

$$H_0: \beta_2 = 3\beta_5$$

Тестовая статистика:

$$LR = -2(l_R(\widehat{\beta_R}) - l_{UR}(\widehat{\beta_{UR}}))\sim\chi^2(1)$$

После реализации LR-теста получаем, что p-value стремится к 0, так как оно меньше любого разумного уровня значимости, то нулевая гипотеза отвергается.

4) Коэффициент при линейной части совпадает по знаку и в k раз больше, чем при нелинейной, а коэффициент при произвольной бинарной переменной равняется t, где $t \neq 0$ можно выбрать произвольным, указав выбранное значение.

Пусть k = 3 и t=1:

$$H_0: \begin{cases} \beta_2 = 3\beta_5 \\ \beta_3 = 1 \end{cases}$$

Тестовая статистика:

$$LR = -2(l_R(\widehat{\beta_R}) - l_{UR}(\widehat{\beta_{UR}})) \sim \chi^2(1)$$

После реализации LR-теста получаем, что p-value стремится к 0, так как оно меньше любого разумного уровня значимости, то нулевая гипотеза отвергается.

Задание 4.5. При помощи LR теста проверьте, можно ли оценивать совместную модель для мужчин и для женщин, либо стоит оценить две различные модели.

$$H_0: \beta_{iMale} = \beta_{iFemale}$$

Тестовая статистика:

$$LR = -2(l_R(\widehat{\beta_R}) - l_{UR}(\widehat{\beta_{UR}})) \sim \chi^2(8)$$

После реализации LR-теста получаем, что p-value стремится к 0, так как оно меньше любого разумного уровня значимости, то нулевая гипотеза отвергается, таким образом, нужно оценивать 2 различные модели для мужчин и женщин.

Задание 4.6*. При помощи LR теста проверьте, можно ли оценивать совместную модель для людей, проживающих в населенных пунктах различного типа (рассмотрите все три возможных типа населенного пункта).

$$H_0: \beta_{iResidence} = \beta_{iVillage} = \beta_{iCity}$$

Тестовая статистика:

$$LR = -2(l_R(\widehat{\beta_R}) - l_{UR}(\widehat{\beta_{UR}})) \sim \chi^2(12)$$

После реализации LR-теста получаем, что p-value стремится к 0, так как оно меньше любого разумного уровня значимости, то нулевая гипотеза отвергается, таким образом, нужно оценивать 3 различные модели для каждого типа населенного пункта.

Часть 5. Логит модель.

Задание 5.1. Оцените логит модель, предварительно записав максимизируемую функцию правдоподобия и указав, чем логит модель отличается от пробит модели. Результат представьте в форме таблицы (можно, например, использовать выдачу из stata, R или python).

$y_i = \begin{cases} 1, & \text{если } y_i^* > 0 \\ 0, & \text{если } y_i^* \le 0 \end{cases}$, где $y_i^* = x_i'\beta + \epsilon_i$ и $\epsilon_i \sim L(0,1)$ (такое распределение ошибок и является отличием логит модели от пробит)

Тогда $P(y_i = 1) = \Lambda(x_i'\beta) = \frac{1}{1+e^{-x_i'\beta}}$

С помощью ММП будут оцениваться коэффициенты модели β:

$$l = \ln(L) = \ln\left(\prod_{i=1}^{N}[\Lambda(x_i'\beta)]^{y_i}\,[1-\Lambda(x_i'\beta)]^{1-y_i}\right)$$
$$= \sum_{i=1}^{N}\left[y_i \ln\left(\Lambda(x_i'\beta)\right) + (1-y_i)\ln\left(1-\Lambda(x_i'\beta)\right)\right] \to \max_{\beta}$$

```
Coefficients:
              Estimate Std. Error z value Pr(>|z|)
(Intercept) -1.752e+00  2.716e-01  -6.450 1.12e-10 ***
age          2.806e-02  8.253e-03   3.400 0.000674 ***
series       1.952e-01  3.030e-02   6.442 1.18e-10 ***
TV          -1.280e+00  6.956e-02 -18.403  < 2e-16 ***
I(age^2)    -1.004e-04  6.406e-05  -1.567 0.117122
age:series  -1.042e-03  4.590e-04  -2.270 0.023233 *
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1
```

Табл. 9. Результаты оценивания логит модели.

Задание 5.2. Проинтерпретируйте значения оценок изменений в отношениях шансов по каждой независимой переменной, входящей линейно.

В модель линейно входит только параметр TV, будем интерпретировать значение оценки изменений в отношениях шансов по ней:

$$\text{Отношение шансов } = \frac{p}{1-p} = e^{x\prime\beta}$$

Тогда изменение отношения шансов:

$$\text{Изменение отношения шансов } = \frac{\mathrm{P(y\ =\ 1 \mid xk\ +\ step)\ /\ P(y\ =\ 0 \mid xk\ +\ step)]}}{\mathrm{P(y\ =\ 1 \mid xk)\ /\ P(y\ =\ 0 \mid xk)}} = e^{\beta}$$

Получаем, что, в случае если индивид начинает часто смотреть телевизор, то отношение шансов, то есть насколько вероятность того, что индвид приобретет подписку, превосходит вероятность того, что индивид не оформит ее, уменьшится на (1 - 0.278) * 100% = 72.2%.

Задание 5.3*. Запишите выражения для расчета изменений в отношениях шансов по каждой независимой переменной, входящей нелинейно. Рассчитайте соответствующие предельные эффекты для индивида с произвольными характеристиками. Результаты расчетов представьте в форме таблицы.

Рассчитывать изменение будем для моих характеристик:

Параметр	age	series	TV	age^2	age*series
Значение	21	10	0	441	210

Для переменной age при увеличении на 1 год:

$$\text{Изменение отношения шансов } = e^{\widehat{\beta_1}+(2age+1)\widehat{\beta_4}+\widehat{\beta_5}series}$$

Для переменной series при увеличении на 1:

$$\text{Изменение отношения шансов } = e^{\widehat{\beta_2}+\widehat{\beta_5}age}$$

Параметр	Изменение отношения шансов
age	1.027008
series	1.414288

Табл. 10. Измененииe в отношениях шансов по независимым переменным age и series.

При увеличении возраста на 1 для индивида с произвольными характеристиками в отношении шансов увеличивается на 2.7%.

При увеличении количества просмотренных за прошлый год сериалов на 1 для индивида с произвольными характеристиками изменение в отношении шансов увеличивается на 41.4%.

Часть 6. Система бинарных уравнений.

Задание 6.1. Оцените систему бинарных уравнений, одно из которых описывает вероятность подписки, а второе — вероятность того, что индивид смотрит телевизор не реже раза в неделю. При этом оба уравнения должны иметь по крайней мере одну общую и одну различающуюся независимую переменную. При необходимости спецификация уравнения подписки может отличаться от той, что использовалась в предыдущих разделах.

Вероятность подписки на онлайн-кинотеатр будет все так же зависеть от возраста, количества просмотренных за год сериалов и факта частого просмотра телевизора.

Вероятность частого просмотра телевизора будет зависеть от возраста, пола и доли свободного времени, проводимого в интернете.

Оцененные уравнения выглядят следующим образом:

```
COPULA:   Gaussian
MARGIN 1: Bernoulli
MARGIN 2: Bernoulli

EQUATION 1
Link function for mu.1: probit
Formula: sub ~ age + series + TV + I(age^2) + age * series

Parametric coefficients:
             Estimate Std. Error z value Pr(>|z|)
(Intercept) -3.155e-01  1.185e-01  -2.664  0.00772 **
age          1.947e-02  3.591e-03   5.421 5.92e-08 ***
series       8.264e-02  1.250e-02   6.610 3.84e-11 ***
TV          -1.954e+00  3.112e-02 -62.778  < 2e-16 ***
I(age^2)    -4.829e-05  2.784e-05  -1.734  0.08284 .
age:series  -3.766e-04  1.940e-04  -1.941  0.05228 .
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1


EQUATION 2
Link function for mu.2: probit
Formula: TV ~ age + male + internet

Parametric coefficients:
             Estimate Std. Error z value Pr(>|z|)
(Intercept)  0.3076615  0.0580315   5.302 1.15e-07 ***
age          0.0135212  0.0008097  16.698  < 2e-16 ***
male        -0.5984762  0.0295116 -20.279  < 2e-16 ***
internet    -1.0687788  0.0626795 -17.051  < 2e-16 ***
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1


n = 5000  theta = 0.957(0.938,0.972)  tau = 0.814(0.774,0.849)
total edf = 11
```

Табл. 11. Результаты оценивания системы бинарных уравнений.

Задание 6.2. Проинтерпретируйте оценки коэффициентов при независимых переменных и коэффициента корреляции между случайными ошибками рассматриваемых уравнений.

Для оцененной модели вероятности оформления подписки на онлайн-кинотеатр: на 5%-ом уровне значимости значимыми оказались линейная часть возраста: с увеличением возраста вероятность оформления подписки увеличивается, количество просмотренных за год сериалов: чем больше сериалов просмотрено, тем больше вероятность подписки, факт частого просмотра телевизора: если человек часто смотрит телевизор, то вероятность оформления им подписки снижается.

Для оцененной модели вероятности частого просмотра телевизора: на 5%-ом уровне значимости значимыми оказались все переменные, так, при увеличении возраста увеличивается вероятность частого просмотра телевизора, если рассматриваемый индивид – мужчина, то вероятность частого просмотра телевизора снижается, чем больше доля свободного времени, проведенного в интернете, тем меньше вероятность частого просмотра телевизора.

Коэффициента корреляции между случайными ошибками рассматриваемых уравнений оказался положительным, что говорит о том, что невключенные переменные одинаково влияют на каждую из рассматриваемых переменных, то есть в обоих случаях включение этих факторов будет одновременно либо увеличивать обе вероятности, либо уменьшать.

Задание 6.3. При помощи LR теста проверьте, имеется ли необходимость в том, чтобы оценивать оба уравнения совместно.

$$H_0: \rho = 0 \text{ (модель, оценивающая отдельно два уравнения)}$$

Тестовая статистика:

$$LR = -2(l_R(\widehat{\beta_R}) - l_{UR}(\widehat{\beta_{UR}})) \sim \chi^2(1)$$

После реализации LR-теста получаем, что p-value стремится к 0, так как оно меньше любого разумного уровня значимости, то нулевая гипотеза отвергается, таким образом, нужно оценивать модели совместно.

Задание 6.4. Для индивида с произвольными характеристиками оцените:

1) Вероятность подписки

Будем оценивать для моих характеристик, но добавим новые параметры (пол – женщина, и доля свободного времени, проводимого в интернете, составляет 0.6):

Параметр	age	series	TV	age^2	age*series	male	internet
Значение	21	10	0	441	210	0	0.6

Получаем, что вероятность подписки составляет 0.79.

2) Вероятность того, что индивид смотрит телевизор по крайней мере раз в неделю

Получаем, что вероятность того, что индивид смотрит телевизор по крайней мере раз в неделю, составялет 0.48.

3) Вероятность того, что индивид и имеет подписку, и смотрит телевизор не реже раза в неделю

Получаем, что вероятность того, что индивид и имеет подписку, и смотрит телевизор не реже раза в неделю, составляет 0.48.

4) Вероятность того, что у индивида имеется подписка, при условии, что он смотрит телевизор реже раза в неделю

Получаем, что вероятность того, что у индивида имеется подписка, при условии, что он смотрит телевизор реже раза в неделю, составляет 0.6.

Часть 7. Сравнение моделей

Задание 7.1. Определите, какая из оцененных вами моделей обладает наибольшей предсказательной силой.

Будем определять предсказательную силу моделей через долю верных предсказаний:

probit	67.30
linprob	67.26
logit	67.30
bp	63.66
naive	63.92

Табл. 12. Доля верных предсказаний пробит, линейно-вероятностной, логит, наивной моделей и системы бинарных уравнений.

Таким образом, видим, что пробит и логит модели показали одинаковое качество, соответственно, каждая из них обладает наибольшей предсказательной силой.

Задание 7.2. Выберите лучшую из оцененных вами моделей руководствуясь информационными критериями.

Для сравнения будем использовать AIC и BIC.

Сначала сравним все одиночные модели друг с другом, а затем лучшую из них сравним с системой бинарных уравнений и, таким образом, выберем лучшую среди всех.

linear	6441.012
probit	6128.620
logit	6129.137

Табл. 13. AIC по линейно-вероятностной, пробит и логит моделей.

linear	6486.633
probit	6167.723
logit	6168.241

Табл. 14. BIC по линейно-вероятностной, пробит и логит моделей.

Так, среди одинарных моделей лучшей оказалась пробит модель. Теперь сравним ее с системой (оценим по отдельности для каждого из уравнений системы пробит модели и просуммируем их критерии):

probit	12049.24
system	11600.79

Табл. 15. AIC по отдельно оценённым пробит моделям для двух уравнений вероятностей и системе.

probit	12114.41
system	11672.48

Табл. 16. BIC по отдельно оценённым пробит моделям для двух уравнений вероятностей и системе.

Получаем, что система бинарных уравнений лучше, чем отдельно оцененные уравнения.