Федеральное государственное автономное образовательное учреждение высшего образования

НАЦИОНАЛЬНЫЙ ИССЛЕДОВАТЕЛЬСКИЙ УНИВЕРСИТЕТ «ВЫСШАЯ ШКОЛА ЭКОНОМИКИ»

Факультет экономических наук

по направлению подготовки Экономика
образовательная программа «Экономика»

Домашняя работа №1

В рамках курса «Микроэконометрика качественных данных»

Выполнил:
Максим Пешков
Группа:
БЭК181

Москва, 2021

Содержание

О данных **2**

Часть 1. Теория и гипотезы **3**
Задание №1.1. . . . 3
Задание №1.2. . . . 4

Часть 2. Линейно-вероятностная модель. **5**
Задание №2.1. . . . 5
Задание №2.2. . . . 6
Задание №2.3. . . . 6

Часть 3. Пробит модель. **8**
Задание №3.1. . . . 8
Задание №3.2. . . . 9
Задание №3.3. . . . 9
Задание №3.4. . . . 10
Задание №3.5. . . . 11
Задание №3.6. . . . 11
Задание №3.7*. . . . 12

Часть 4. Тестирование корректности спецификации пробит модели. **12**
Задание №4.1. . . . 12
Задание №4.2. . . . 13
Задание №4.3. . . . 14
Задание №4.4. . . . 14
Задание №4.5. . . . 16
Задание №4.6*. . . . 16

Часть 5. Логит модель. **17**
Задание №5.1. . . . 17
Задание №5.2. . . . 18
Задание №5.3*. . . . 18

Часть 6. Система бинарных уравнений. **19**
Задание №6.1. . . . 19
Задание №6.2. . . . 19
Задание №6.3. . . . 20
Задание №6.4. . . . 20

Часть 7. Сравнение моделей **22**
Задание №7.1. . . . 22
Задание №7.2. . . . 22

Часть 8. Модель бинарного выбора со случайными ошибками, имеющими распределение Стьюдента **24**
Задание №8.1.*** . . . 24
Задание №8.2.*** . . . 24
Задание №8.3.*** . . . 25

Список таблиц

1 Результаты линейно-вероятностной модели 5
2 Средние предельные эффекты для линейно-вероятностной модели 7
3 Результаты пробит модели 8
4 Характеристики рассматриваемого индивида 10
5 Средние предельные эффекты для пробит модели 11
6 Доли верных предсказаний пробит, линейной-вероятностной, наивной модели . . 11
7 Результаты логит модели 17
8 Выражения для расчета изменений в отношениях шансов по каждой независимой переменной, входящей нелинейно и их значения 18
9 Характеристики рассматриваемого индивида в системах уравнений 21
10 Доли верных предсказаний пробит, линейной-вероятностной, наивной модели, логит модели и модели системы бинарных уравнений 22
11 Общие информационные критерии AIC, BIC для пробит, линейной-вероятностной, наивной модели, логит модели и модели системы бинарных уравнений 22
12 Информационные критерии AIC, BIC для пробит, линейной-вероятностной, наивной модели, логит модели и модели системы бинарных уравнений для 2 выбранных уравнений 23
13 Оценки бинарной модели со случайными ошибками, имеющими распределение Стьюдента с 13 степенями свободы 25

О данных

В данной работе мы будем изучать, как различные факторы влияют на вероятность того, что индивид оформит подписку на онлайн кинотеатр.

Данные содержат информацию о следующих индивидуальных характеристиках:

- income — доход
- age — возраст
- internet — доля свободного времени, проводимого в интернете
- series — количество просмотренных за год сериалов
- health — субъективная оценка здоровья
- male — половая принадлежность (1 – мужчина, 0 – женщина)
- marriage — состоит в официальном браке
- residence — место проживания
- cat — факт наличия кота
- news — субъективная оценка степени, в которой индивид интересуется новостями
- sub — факт наличия подписки на онлайн кинотеатр (**зависимая переменная**)
- TV — индивид смотрит телевизор не реже раза в неделю.

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.

В качестве независимых переменных будут использоваться переменные **internet, TV, age, series**.

- **series**

  На выбор покупки онлайн подписки влияет заинтересованность человека в просмотре сериалов в целом. И чем больше индивид смотрит сериалов, тем больше вероятность, что он использует онлайн кинотеатр, а если он их смотрит мало, то такие издержки за оплату подписки ему будут невыгодны и он не будет использовать онлайн кинотеатр вовсе. Поэтому ожидается наблюдение положительного эффекта.

- **internet**

  Чем чаще люди проводят время в интернете, тем более осведомлены о подписках в онлайн-кинотеатрах, а также им может казаться проще и выгоднее оформить подписку, так как на сегодняшний день большая часть онлайн кинотеатров существуют в одной экосистеме с другими сервисами (например, Кинопоиск и Яндекс). Поэтому ожидается наблюдение положительного эффекта.

- **TV**

  Такие способы просмотра сериалов или кино как телевидение и онлайн-кинотеатры могут быть взаимозаменяемыми для некоторых людей, поэтому выбор одного может исключать выбор другого, так как при покупке подписки уже не будет смысла в телевидении (на платформе онлайн-кинотеатра часто есть те же самые фильмы или вовсе уникальные), а при частом просмотре телевидения может быть невыгодно переходить на онлайн-кинотеатр, так как интересные для индивида фильмы показывают по телевидению. Поэтому ожидается наблюдение отрицательного эффекта.

- **age**

  Выбор этой переменной обоснован тем, что заинтересованность в онлайн-кинотеатрах различна среди поколений. Старшее поколение и более взрослые люди воспринимают это как новинку и не видят в нем необходимость, так как часто используют телевидение для просмотра фильмов или сериалов. А молодое поколение наоборот, чаще использует интернет и онлайн-просмотр для фильмов. Поэтому ожидается наблюдение положительного эффекта.

Задание №1.2.

Сформулируйте по крайней мере одну гипотезу о наличии эффекта взаимодействия и 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.

Будут проверяться две нелинейные гипотезы: о наличии эффекта взаимодействия переменных **internet, TV**; о наличии квадратичного эффекта **age**. Ниже приведены обоснования для этого.

- **Эффекта взаимодействия переменных internet, TV**

  Использование интернета по-разному влияет на выбор подписки у людей, которые часто смотрят телевидение и нет. Чем чаще индивид смотрит телевидение, тем для его выбора частое использование интернета менее важно, так как, скорей всего, он просматривает интересующие его фильмы или сериалы по ТВ и у него нет заинтересованности в оформлении подписки на онлайн кинотеатр. А при не частом просмотре телевидения большое количество времени в интернете может как раз и означать, что он тратит его часть на просмотры чего-либо в онлайн кинотеатре. Тем самым, ожидается наблюдение отрицательного эффекта у интеракции между переменными - отрицательный эффект взаимодействия.

- **Квадратичная зависимость от age**

  Оплата подписки в онлайн кинотеатре может казаться высокой для людей с низким уровнем дохода, которыми чаще всего являются молодые люди или пенсионеры. Более того, у молодых людей может не хватать свободного времени на просмотры фильмов или сериалов в любом формате из-за других интересов в свободное время: активная учеба, занятия спортом или прогулки и общения с друзьями. А для более старого поколения использование онлайн-кинотеатров может быть сложным и труднодоступным. Поэтому ожидается наблюдение перевернутого U-образного эффекта от возраста, то есть перед линейной компонентой будет положительный знак, а перед квадратичной - отрицательный (отрицательный эффект для молодых и пожилых людей).

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).

Для проверки выше перечисленных гипотез в линейно-вероятностной модели будет использоваться следующее регрессионное уравнение:

$$sub_i = \beta_0 + \beta_1 \times series_i + \beta_2 \times internet_i + \beta_3 \times TV_i + \beta_4 \times age_i + \beta_5 \times age_i^2 + \beta_6 \times internet_i \times TV_i + \varepsilon_i$$

Для нахождения оцениваемых параметров β_i в линейно-вероятностной модели используется метод наименьших квадратов (МНК), то есть параметры оцениваются и находятся аналитически по такой формуле $\hat{\beta} = (X^T X)^{-1} X^T$. Результаты найденных оценок представлены ниже в Таблице 1.

Таблица 1: Результаты линейно-вероятностной модели

	Dependent variable:
	sub
series	0.026^{***}
	(0.002)
internet	0.248^{***}
	(0.044)
TV	-0.227^{***}
	(0.030)
age	0.005^{***}
	(0.002)
age^2	-0.00002^{*}
	(0.00001)
internet×TV	-0.045
	(0.054)
Constant	0.096^{*}
	(0.052)
Observations	5,000
R^2	0.090
Adjusted R^2	0.089
F Statistic	82.296^{***} (df = 6; 4993)
AIC	6365.68
BIC	6417.82
Note:	*p<0.1; **p<0.05; ***p<0.01

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.

Недостатки линейно-вероятностной модели заключаются в следующем:

- Гетероскедастичность

  Из-за того, что $Var(\varepsilon_i) = x_i'\beta(1 - x_i'\beta)$, то есть для каждого наблюдения ошибка зависит от зависимых переменных, то в этой модели есть гетероскедастичность. Следствием чего является неэффективность оценок.

- Ненормальность распределения ошибок

  В этой модели ошибки распределены следующим образом:

  $$\begin{cases} \varepsilon_i = 1 - x_i'\beta, \text{ если } y_i = 1 \\ \varepsilon_i = -x_i'\beta, \text{ если } y_i = 0 \end{cases}$$

  Это ведет к тому, что не следует проверять гипотезы и интерпретировать полученные R^2 и F-статистику при малом количество наблюдений.

- Неинтерпретируемость оценок коэффициентов

  Через оценки данной модели выражается вероятность $\hat{p}_i = x_i'\hat{\beta}$, которая может не лежать на интервале $[0, 1]$, так как при оценки МНК нет ограничения на вхождение в этот интервал, из-за чего невозможно правильно интерпретировать оценки коэффициентов.

Предложения для оценки качества модели и проверки гипотез:

- Для борьбы с гетероскедастичностью можно использовать Обобщенный МНК с ковариационной матрицей устойчивой к гетероскедастичности
- Для оценивания качества модели можно использовать AIC, BIC или качество точности модели - *accuracy*
- Для оценивания гипотез о значимости коэффициентов можно использовать бутстрапированные доверительные интервалы и проверять будет ли лежать 0 в данном ассимптотическом доверительном интервале

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.

Формулы для оценивания предельных эффектов в этой модели:

$$\frac{\partial sub_i}{\partial series_i} = \hat{\beta}_1$$
$$\frac{\partial sub_i}{\partial internet_i} = \hat{\beta}_2 + \hat{\beta}_6 * TV_i$$
$$\frac{\partial sub_i}{\partial age_i} = \hat{\beta}_5 + 2 * \hat{\beta}_6 * age_i$$
$$sub_i|_{TV_i=1} - sub_i|_{TV_i=0} = \beta_3 + \beta_6 * internet_i$$

Средний предельный эффект может оцениваться 2 способами: среднее всех предельных эффектов в точках (AME), предельный эффект в точке со средними показателями (MME). Так как в условиях не сказано какой из них считать, то будем расчитать MME.

Ниже представлена Таблица 2 с найденными средними предельными эффектами для переменных $age, internet, TV$, а также предельный эффект для $series$, записанные в колонке MME. Также для переменных входящих нелинейно записано когда предельный эффект принимает отрицательные значения, а когда положительные.

Таблица 2: Средние предельные эффекты для линейно-вероятностной модели

переменная	MME	отрицательные	положительные	min	max
series	0.0262	-	-	0.0262	0.0262
internet	0.2190	-	$\forall$ TV	0.212	0.2475
TV	-0.2472	при любых значениях internet	-	-0.2677	-0.2272
age	0.0017	[95.68114, +inf)	[0, 95.68114)	-0.0002	0.0036

Как видно из таблицы предельные эффекты для $internet, series$ положительные, что говорит о положительном направлении эффектов от количества просмотренных сериалов и времени в интернете на оформление подписки онлайн-кинотеатра. Средний предельный эффект от количества времени в интернете говорит о том, что при увеличении доли свободного времени в интернете на 10% (0.1) вероятность оформления подписки вырастет на 21.9% при прочих равных, а вот увеличение просмотренных сериалов на 1 увеличивает вероятность оформления подписки на 26.2% при прочих равных.

Однако разница в частоте просмотра телевидения дает негативный предельный эффект, как и ожидалось, то есть при увеличении частоты просмотра ТВ и изменения переменной с 0 до 1 вероятность уменьшается на 24.72%.

Для возраста предельный эффект разный в зависимости от возраста. Для людей младше 95 лет предельный эффект положительный, а для людей старше отрицательный.

Часть 3. Пробит модель.

Задание №3.1.

Оцените пробит модель, предварительно записав максимизируемую функцию правдоподобия, 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).

Для оценивания пробит модели применяется метод максимального правдоподобия, где максимизируемая функцию правдоподобия равна

$$Ł = \prod_{i=1}^{n} \left(\mathcal{F}(x_i'\beta)\right)^{sub_i} \left(1 - \mathcal{F}(x_i'\beta)\right)^{1-sub_i} \to max_{\beta_i}$$

Где x_i - вектор $(series_i, internet_i, TV_i, age_i, age_i^2, internet_i \times TV_i)$, β - тоже вектор оцениваемых параметров $\beta_i, i = \overline{0,6}$, $\mathcal{F} = \frac{1}{\sqrt{2\pi}} \int_{-\inf}^{z} e^{\frac{t^2}{2}} dt$ - функция стандартного нормального распределения.

Результаты найденных оценок представлены в Таблице 3.

Таблица 3: Результаты пробит модели

	Dependent variable:
	sub
series	0.077^{***}
	(0.007)
internet	0.632^{***}
	(0.123)
TV	-0.693^{***}
	(0.087)
age	0.013^{***}
	(0.005)
age^2	-0.0001^{*}
	(0.00004)
internet×TV	-0.007
	(0.154)
Constant	-1.135^{***}
	(0.151)
Observations	5,000
Log Likelihood	−3,021.975
Akaike Inf. Crit.	6,057.949
Note:	*p<0.1; **p<0.05; ***p<0.01

Основные свойства оценок пробит модели, полученных методом максимизации правдоподобия, заключаются в

- Состоятельны
- Асимптотически несмещенные
- Асимптотически нормальны
- Асимптотически эффективны
- Инвариантны

Задание №3.2.

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

Для пробит модели само значение коэффициента нельзя никак численно интерпретировать, можно только его знак для некоторых переменных, что и будет сделано ниже.

- Для переменной *series* коэффициент положительно значим для вероятности оформления подписки онлайн кинотеатра. Такой же положительный эффект для количества просмотренных сериалов предполагался выше.
- Частота просмотра телевидения TV отрицательно и значимо влияет на вероятность оформления подписки онлайн кинотеатра, что соотносится с высказанными предположениями
- Доля времени в интернете *internet* положительно и значимо влияет на вероятность оформление подписки онлайн кинотеатра, что соотносится с высказанными предположениями
- Возраст *age* влияет на вероятность наличия подписки онлайн кинотеатра перевернутой U-образной зависимостью, как и ожидалось в выказанных предположениях. Хотя и коэффициент перед линейной переменной значим на любом адекватном уровне значимости, а перед квадратом только на 10 % значимости.
- По оцененным параметрам можно сказать, что не наблюдается значимости перед выбранным эффектом взаимодействия, причем ожидалась отрицательная значимость. Хотя оценка коэффициента перед интеракцией доли времени в интернете и частотой просмотра телевидения имеет отрицательный знак, что соответствует о предположении направления эффекта.

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).

Оценим вероятность наличия подписки для индивида с такими характеристиками, представленные в Таблице 4.

Таблица 4: Характеристики рассматриваемого индивида

Переменная	Значение
$series_{ind}$	4
$internet_{ind}$	0.7
TV_{ind}	0
age_{ind}	21

Вероятность наличия подписки для индивида оценивается по следующей формуле:

$$P(sub_{ind}=1) = \mathcal{F}(\hat{\beta}_0 + \hat{\beta}_1 \times series_{ind} + \hat{\beta}_2 \times internet_{ind} + \hat{\beta}_3 \times TV_{ind} \\ + \hat{\beta}_4 \times age_{ind} + \hat{\beta}_5 \times age_{ind}^2 + \hat{\beta}_6 \times internet_{ind} \times TV_{ind})$$

Где $\hat{\beta}_i$ - оцененные параметры в модели, вместо регрессоров подставляются их соответствующие значения, а $\mathcal{F} = \frac{1}{\sqrt{2\pi}} \int_{-\inf}^{z} e^{\frac{t^2}{2}} dt$ - функция стандартного нормального распределения.

$$P(sub_{ind}=1) = \mathcal{F}(-1.135 + 0.077 \times series_{ind} + 0.632 \times internet_{ind} - 0.693 \times TV_{ind} \\ + 0.013 \times age_{ind} - 0.0001 \times age_{ind}^2 - 0.007 \times internet_{ind} \times TV_{ind})$$
$$P(sub_{ind}=1) = \mathcal{F}(-1.135 + 0.077 \times 4 + 0.632 \times 0.7 - 0.693 \times 0 + 0.013 \times 21 - 0.0001 \times 21^2 \\ - 0.007 \times 0.7 \times 0)$$

Найденное значение равно 0.4464257, то есть с вероятностью в 44.64% у рассмотренного индивида есть подписка на онлайн кинотеатр.

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.

Оценим средние предельные эффекты для непрерывной переменной *age* и бинарной переменной TV. Формулы для нахождения их среднего предельнего эффекты (MME):

- для непрерывной переменной *age* средний предельный эффект считается как

$$\frac{\partial P(sub=1)}{\partial age} = f(\overline{x}'\hat{\beta}) \times (\hat{\beta}_4 + 2 * \hat{\beta}_6 * \overline{age})$$

  Где $\overline{x}$ - средние значения регрессоров модели, $\overline{age}$ - среднее значение возраста в выборке, а $f = \frac{1}{\sqrt{2\pi}} e^{\frac{t^2}{2}}$ - функция плотности стандартного нормального распределения.

- для бинарной переменной TV средний предельный эффект считается как разница

$$\Delta P(sub=1) = P(sub=1|\overline{x_{-TV}}, TV=1) - P(sub=1|\overline{x_{-TV}}, TV=0) = \\ = \mathcal{F}(\hat{\beta}_0 + \hat{\beta}_1 \times \overline{series} + \hat{\beta}_2 \times \overline{internet} + \hat{\beta}_3 \times 1 + \hat{\beta}_4 \times \overline{age} + \hat{\beta}_5 \times \overline{age}^2 + \hat{\beta}_6 \times \overline{internet} \times 1) - \\ - \mathcal{F}(\hat{\beta}_0 + \hat{\beta}_1 \times \overline{series} + \hat{\beta}_2 \times \overline{internet} + \hat{\beta}_3 \times 0 + \hat{\beta}_4 \times \overline{age} + \hat{\beta}_5 \times \overline{age}^2 + \hat{\beta}_6 \times \overline{internet} \times 0)$$

Где $\overline{x_{-TV}}$ - средние значения регрессоров модели кроме бинарной переменной TV, а $f = \frac{1}{\sqrt{2\pi}} e^{\frac{t^2}{2}}$ - функция плотности стандартного нормального распределения

Результаты расчета средних предельных эффектов представлены в Таблице 5.

Таблица 5: Средние предельные эффекты для пробит модели

Переменная	MME
age	0.001829614
TV	-0.2631863

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.

Посчитав доли верных предсказаний, представленных в Таблице 6, можно сказать, что предсказательная сила пробит модели лучше наивной модели, но хуже чем у линейной-вероятностной модели. Тем самым, можно заключить, что пробит модель хотя и позволяет получить интерпретируемые результаты, но имеет не самую лучшую предсказательную силу.

Таблица 6: Доли верных предсказаний пробит, линейной-вероятностной, наивной модели

Модель	Доля верных предсказаний
Пробит модель	0.6760
Наивная модель	0.6446
Линейно-вероятностная модель	0.6774

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.

Проверим гипотезу о том, что предельный эффект на вероятность наличия подписки по переменной *series* является значимым для индивида с произвольными характеристиками, при этом зная формулу для предельного эффекта *series* запишем нулевую и альтернативную гипотезы.

$$\begin{cases} H_0 : \dfrac{\partial P(sub_i = 1)}{\partial series_i} = f(x'\hat{\beta}) \times \hat{\beta}_1 = 0 \\ H_a : \dfrac{\partial P(sub_i = 1)}{\partial series_i} = f(x'\hat{\beta}) \times \hat{\beta}_1 \neq 0 \end{cases}$$

Для проверки этой гипотезы необходимо использовать дельта-метод (чтобы найти дисперсию предельного эффекта), после которого тестовая статистика имеет z-распределение, так как распределение тестовой статистики будет асимптотически нормальным.

Так как в условии задания не написано, что необходимо проверить гипотезу "руками то воспользуемся пакетом *margins*, которые не только считает средние предельные эффекты, но и считает z-статистику и p-value для проверки гипотезы о значимости предельного эффекта.

Используя пакет *margins*, легко получить, что z -статистика равна 5.8802, p-value=0.000, то есть нулевая гипотеза отвергается и предельный эффект *series* является значимым для индивида с произвольными характеристиками.

Задание №3.7*.

Повторите предыдущий пункт для переменной, имеющей взаимодействие.

Проверим гипотезу о том, что предельный эффект на вероятность наличия подписки по переменной *internet* является значимым для индивида с произвольными характеристиками, при этом зная формулу для предельного эффекта *internet* запишем нулевую и альтернативную гипотезы.

$$\begin{cases} H_0 : \dfrac{\partial P(sub_i = 1)}{\partial internet_i} = f(x'\hat{\beta}) \times (\hat{\beta}_2 + \hat{\beta}_7 \times TV) = 0 \\ H_a : \dfrac{\partial P(sub_i = 1)}{\partial internet_i} = f(x'\hat{\beta}) \times (\hat{\beta}_2 + \hat{\beta}_7 \times TV) \neq 0 \end{cases}$$

Для проверки этой гипотезы необходимо использовать дельта-метод (чтобы найти дисперсию предельного эффекта), после которого получится z-статистика, так как распределение тестовой статистики будет асимптотически нормальным.

Так как в условии задания не написано, что необходимо проверить гипотезу "руками то воспользуемся пакетом *margins*, которые не только считает средние предельные эффекты, но и считает z-статистику и p-value для проверки гипотезы о значимости предельного эффекта.

Используя пакет *margins*, легко получить, что z -статистика равна 8.5805, p-value=0.000, то есть нулевая гипотеза отвергается и предельный эффект *internet* является значимым для индивида с произвольными характеристиками.

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.

Нарушение допущения о нормальности распределения ошибок может привести к таким негативным последствиям как

- Несостоятельность оценок
- Возможные искажения при оценивание значимости коэффициентов
- Неверность проводимых тестов на значимость

Для проверки нормальности распределения ошибок воспользуемся LM тестом с использованием регрессии на единичный вектор. А именно благодаря найденным обобщенным остаткам находим градиенты по оцениваемых параметров и, регрессируя единичный вектор на них, находим R^2. Тогда статистика для LM теста $R^2 * n$ также будет совпадать с методом через аналитические функции и будет иметь такую же хи квадрат статистику с 2 степенями свободами.

В этом тесте нулевая гипотеза будет заключаться в наличие нормального распределения ошибок, а альтернативная в отсутствии нормального распределения.

$$\begin{cases} H_0 : \varepsilon_i \sim \mathcal{N}(0, \sigma) \\ H_a : \varepsilon_i \nsim \mathcal{N}(0, \sigma) \end{cases}$$

Проведя LM тест, получили статистику равную 0.105, а p-value равную 0.9485. То есть нулевая гипотеза не отвергается, то есть после построения регрессии наблюдается нормальность распределения ошибок, что в целом было ожидаемо, так как в выборке достаточно большое количество наблюдений (5000), а при большом количестве наблюдений по ЦПТ распределение стремится к нормальному.

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.

Как мы видели в результатах пробит модели в таблице 3, то коэффициент перед интеракцией не значим, хотя мы предполагали значимость, что может быть вызвано гетероскедастичностью. Будем предполагать, что на дисперсию случайной ошибки влияют такие переменные, как $series, income$, так как

- *series* входит в линейный индекс основного уравнения

  Влияние количества сериалов может влиять на дисперсию ошибки, ведь при просмотре до 3 сериалов может быть такое же безразличие и незаинтересованность в оформлении подписки на онлайн кинотеатр, а при просмотре 9 и больше сериалов появляется "сериальная"зависимость и соотношение эффектов на вероятности оформления подписки будет больше. Поэтому предполагается, что возможно наличие гетероскедастичности с зависимостью от *series*.

- *income* входит в линейный индекс дисперсии случайной ошибки

  Влияние дохода очевидно при оформлении подписки, ведь не каждый человек может позволить себе тратить ежемесячно 200-300 рублей, например, достаточно бедные люди не будут заинтересованы в этом. Поэтому ошибки могут коррелировать с уровнем дохода, что ведет к гетероскедастичности.

Гетероскедастичность и нарушение допущения о гомоскедастичности может привести к таким негативным последствиям как:

- Неэффективность оценок
- Нарушение предположений ТГМ могут вести и к неправильной инерпретации значимости коэффициентов
- Возможно искажение в качестве модели и уменьшения предсказательной силы

Для проверки наличия гетероскедастичности воспользуемся LR тестом. Нулевая гипотеза заключается в наличии гомоскедастичности ошибок, а альтернативная в гетероскедастичности.

$$\begin{cases} H_0 : \varepsilon_i \sim \mathcal{N}(0, \sigma) \\ H_a : \varepsilon_i \sim \mathcal{N}(0, \sigma_i) \end{cases}$$

А если более точнее, то альтернативная гипотеза предполагает, что дисперсия зависима от $series, income$ как $\sigma_i = \exp(series_i * \tau_{1i} + income_i * \tau_{2i})$, а при нулевой предполагаем, что все τ равны 0.

$$\begin{cases} H_0 : \tau = 0 \\ H_a : \text{хотя бы один из } \tau \text{ не равен } 0 \end{cases}$$

Тестовая статистика равна 56.438 и имеет хи квадрат распределение с 2 степенями свободы, а p-value равен 5.554e-13, то есть нулевая гипотеза отвергается на любом разумном уровне значимости. Откуда следует, что ошибки модели гетероскедастичны.

Однако выводы могут быть не совсем корректны,так как LR тест использует явный вид функции h для линейного индекса дисперсии, хотя при помощи LM теста нет необходимости предполагать конкретную форму функции h достаточно наложить на нее лишь пару ограничений: $h(0) = 1, h'(0) \neq 0$. Поэтому здесь следует еще провести LM тест, чтобы убедиться в том, что нет зависимости от выбора вида функции h и ошибки модели действительно гетероскедастичны.

Задание №4.3.

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

Для модели с гетероскедастичной случайной ошибкой запишем основные уравнения:

$$\begin{aligned} P(sub = 1) = \mathcal{F}(&\hat{\beta}_0 + \hat{\beta}_1 \times series + \hat{\beta}_2 \times internet + \hat{\beta}_3 \times TV \\ &+ \hat{\beta}_4 \times age + \hat{\beta}_5 \times age^2 + \hat{\beta}_6 \times internet \times TV) \end{aligned}$$

$$\sigma_i = \exp(series_i * \tau_{1i} + income_i * \tau_{2i})$$

Как мы знаем из лекций и семинаров, то путем нехитрых операций дифференцирования можно получить формулу для вычисления среднего предельного эффекта:

$$\frac{\partial P(sub = 1)}{\partial series} = f(\overline{x}'\widetilde{\beta_{het}}, sd = \widetilde{\sigma_{het}}) \times (\widetilde{\beta_{het}^{series}} - \widetilde{\tau_1} \times \overline{series})$$

Где $\widetilde{\beta_{het}}, \widetilde{\tau_1}, \widetilde{\tau_2}, \widetilde{\sigma_{het}}$ - оценки параметров пробит модели с учетом гетероскедастичности при значениях $\overline{x}$, f - функция плотности нормального распределения с матожиданием 0 и дисперсией $\widetilde{\sigma_{het}}$.

Подставляя найденные значения получаем, что предельный эффект на вероятность равен 0.02641532 (что совпадает с тем, если считать производную численно).

А предельный эффект на дисперсию случайной ошибки считается проще:

$$\frac{\partial \sigma}{\partial series} = 2 \times \widetilde{\tau_1} \times \exp(\overline{series} * \widetilde{\tau_1} + \overline{income} * \widetilde{\tau_2})$$

Подставляя найденные значения получаем, что предельный эффект на дисперсию случайной ошибки равен 0.04119267.

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*
2. *Оба коэффициента равняются нулю*
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.*
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.*

Проведем LR тесты для переменной *age* по очереди:

1. Коэффициент при линейной части β_4 равняется нулю

$$\begin{cases} H_0 : \beta_4 = 0 \\ H_a : \beta_4 \neq 0 \end{cases}$$

   Ограниченная модель - когда коэффициент β_4 равен 0, а полная -обычная пробит модель.

   Полученное тестовое значение из хи квадрат распределения с 1 степенью свободы равно 7.5546, а p-value = 0.005986, то есть нулевая гипотеза отвергается на 1% уровне значимости. То есть коэффициент не равен нулю.

2. Оба коэффициента β_4, β_5 равняются нулю

$$\begin{cases} H_0 : \beta_4, \beta_5 = 0 \\ H_a : \beta_4 \neq 0 \text{ или } \beta_5 \neq 0 \end{cases}$$

   Ограниченная модель - когда коэффициенты β_4, β_5 равны 0, а полная - обычная пробит модель.

   Полученное тестовое значение из хи квадрат распределения с 2 степенями свободы равно 35.854, а p-value = 1.639e-08, то есть нулевая гипотеза отвергается на 1% уровне значимости. То есть коэффициенты не равны нулю.

3. Коэффициент β_4 при линейной части совпадает по знаку и в 10 раз больше (по модулю), чем при нелинейной β_5, то есть $\beta_4 = 10 * \beta_5$

$$\begin{cases} H_0 : \beta_4 - 10 * \beta_5 = 0 \\ H_a : \beta_4 - 10 * \beta_5 \neq 0 \end{cases}$$

   Ограниченная модель - когда коэффициент $\beta_4 = 10 * \beta_5$, а полная - обычная пробит модель.

   Полученное значение тестовой статистики из распределения хи квадрат с 1 степенями свободы равно 7.5504, а p-value = 0.006, то есть нулевая гипотеза отвергается на 1% уровне значимости.

4. Коэффициент при линейной части совпадает по знаку и в 10 раз больше, чем при нелинейной, а коэффициент β_3 при бинарной переменной TV равняется -0.5. Этот пункт делается через offset.

$$\begin{cases} H_0 : \beta_4 - 10 * \beta_5 = 0, \beta_3 = 0.5 \\ H_a : \beta_4 - 10 * \beta_5 \neq 0 \text{ или } \beta_3 \neq 0.5 \end{cases}$$

Ограниченная модель - когда коэффициент $\beta_4 = 10 * \beta_5$ и $\beta_3 = 0.5$, а полная - обычная пробит модель.

Полученное тестовое значение из хи квадрат распределения с 2 степенями свободы равно 12.033, а p-value = 0.002438, то есть нулевая гипотеза отвергается на 1% уровне значимости.

Задание №4.5.

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

В этом случае ограниченная модель будет иметь следующий вид (такая же как и обычная, только с включением переменной пола)

$$P(sub_i = 1) = \mathcal{F}(\hat{\beta}_0 + \hat{\beta}_1 \times series_i + \hat{\beta}_2 \times internet_i + \hat{\beta}_3 \times TV_i + \hat{\beta}_4 \times age_i + \hat{\beta}_5 \times age_i^2 + \hat{\beta}_6 \times internet_i \times TV_i) + \hat{\beta}_7 \times male_i$$

А полную модель представим как комбинацию двух пробит-моделей с изначальной формулой, только на отдельных данных по женщинам и мужчинам.

$$\begin{cases} H_0 : \text{коэффициенты в моделях не различаются} \\ H_a : \text{коэффициенты в моделях различаются} \end{cases}$$

Тестовая статистика равна 24.96 и имеет распределение хи квадрат с 6 степенями свободы, p-value=0.0003470173, то есть нулевая гипотеза отвергается на уровне значимости на уровне значимости 1%. Следовательно, нужно оценивать две различные модели для мужчин и для 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).

В этом случае ограниченная модель будет иметь следующий вид (такая же как и обычная, только с включением дамми на типы населенного пункта: $[residence_i = City], [residence_i = Capital]$)

$$P(sub = 1) = \mathcal{F}(\hat{\beta}_0 + \hat{\beta}_1 \times series_i + \hat{\beta}_2 \times internet_i + \hat{\beta}_3 \times TV_i + \hat{\beta}_4 \times age + \hat{\beta}_5 \times age_i^2 + \hat{\beta}_6 \times internet_i \times TV_i) + \hat{\beta}_7 \times [residence_i = City] + \hat{\beta}_8 \times [residence_i = Capital]$$

А полную модель представим как комбинацию трех пробит-моделей с изначальной формулой, только на отдельных данных по населенным пунктам.

$$\begin{cases} H_0 : \text{коэффициенты в моделях не различаются} \\ H_a : \text{коэффициенты в моделях различаются} \end{cases}$$

Тестовая статистика равна 12.4 и имеет распределение хи квадрат с 7 степенями свободы, p-value=0.08810467, то есть нулевая гипотеза отвергается на уровне значимости 10%, но не отвергается на уровне значимости 5%. Следовательно, нужно оценивать совместную модель для населенных пунктов.

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).

Для оценивания логит модели применяется метод максимального правдоподобия, где максимизируемая функцию правдоподобия равна

$$Ł = \prod_{i=1}^{n} \left(\Lambda(x_i'\beta)\right)^{sub_i} \left(1 - \Lambda(x_i'\beta)\right)^{1-sub_i} \to max_{\beta_i}$$

Где x_i - вектор $(series_i, internet_i, TV_i, age_i, age_i^2, internet_i \times TV_i)$, β - тоже вектор оцениваемых параметров $\beta_i, i = \overline{0,6}$, $\Lambda(z) = \frac{1}{1+e^{-z}}$ - функция распределения логистического распределения. Отличие логит модели от пробит заключается в предположении другого распределения случайных ошибок - логистическое, поэтому и в функции правдоподобия используется логистическая функция распределения.

Результаты найденных оценок представлены в Таблице 7.

Таблица 7: Результаты логит модели

	Dependent variable:
	sub
series	0.127^{***}
	(0.011)
internet	1.012^{***}
	(0.200)
TV	-1.162^{***}
	(0.144)
age	0.022^{***}
	(0.008)
age^2	-0.0001^{*}
	(0.0001)
internet × TV	0.034
	(0.254)
Constant	-1.860^{***}
	(0.251)
Observations	5,000
Log Likelihood	−3,021.943
Akaike Inf. Crit.	6,057.887
Note:	*p<0.1; **p<0.05; ***p<0.01

Задание №5.2.

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

В моей модели только одна независимая переменная *series*, входящая линейно. Для нее и посчитаем значения оценок изменений в отношениях шансов, где отношение шансов $\frac{P(sub_i = 1)}{P(sub_i = 0)}$, то есть отношение вероятностей, что есть оформленная подписка на онлайн кинотеатр к той, когда ее нет.

Формулы отношение шансов и оценки их изменения по переменной *series* для логит модели будут выглядеть так

$$OR_i = \frac{P(sub_i = 1)}{P(sub_i = 0)} = \frac{\Lambda(x_i'\beta)}{1 - \Lambda(x_i'\beta)} = e^{x_i'\beta}$$
$$\frac{OR_2}{OR_1} = e^{\beta_1 \times (series_2 - series_1)}$$

Различие между 2 и 1 состоянием образно оценим в 1, чтобы получить приращение отношения шансов при росте *series* на 1 (выбран именно этот шаг, так как *series* целое). Тогда оценка изменений в отношениях шансов равна 1.135362. Это означает, что при увеличении количества просмотренных на 1 отношение вероятностей, что есть оформленная подписка на онлайн кинотеатр к той, когда ее нет, увеличивается в 1.135362 раз.

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.

Выражения для расчета изменений в отношениях шансов по каждой независимой переменной, входящей нелинейно $age, internet, TV$ и их значения для индивида, рассматриваемый ранее, представлены в Таблице 8. Там же представлены значения шаг для каждой переменной (step): различие между состояниями изменения (для предыдущего пункта он был равен 1), которые выбраны исходя из распределения переменных.

Напомню, что характеристики выбранного индивида такие же и представлены в Таблице 4.

Таблица 8: Выражения для расчета изменений в отношениях шансов по каждой независимой переменной, входящей нелинейно и их значения

Переменная	Выражения для расчета	Шаг (step)	Значение
age	$\exp(\beta_4 * step + \beta_5 * (2 * step * age_{ind} + step^2))$	1	1.017316
internet	$\exp(\beta_2 * step + \beta_6 * step * TV_{ind})$	0.1	1.106496
TV	$\exp(\beta_3 * step + \beta_6 * step * series_{ind})$	1	0.8923633

Из Таблицы 8 видно, что

- При увеличении возраста на 1 отношение вероятностей, что есть оформленная подписка на онлайн кинотеатр к той, когда ее нет, увеличивается в 1.017316 раз.

- При увеличении доли свободного времени на 0.1 отношение вероятностей, что есть оформленная подписка на онлайн кинотеатр к той, когда ее нет, увеличивается в 1.106496 раз.
- При увеличении частоты пользования телевидения (переходе в состояние частых пользователей) отношение вероятностей, что есть оформленная подписка на онлайн кинотеатр к той, когда ее нет, уменьшается в 0.8923633 раз.

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.

Зададим следующую спецификацию системы бинарных уравнений, где общей независимой переменной будет *internet*, а случайные ошибки имеют совместное нормальное стандартное распределение:

$$\begin{cases} P(sub_i) = \beta_0 + \beta_1 \times series_i + \beta_2 \times internet_i + \beta_3 \times income_i + \varepsilon_i \\ P(TV_i) = \gamma_0 + \gamma_1 internet_i + \gamma_2 age_i + u_i \end{cases}$$

Выбор независимых переменных для TV основывался на том, что если человек часто пользуется интернетом, то скорей всего в меньшей мере пользуется телевидением, а также более взрослые люди чаще смотрят телевидение. Также было решение добавить доход как независимую переменную, так как более богатые люди могут себе позволить подписку на онлайн кинотеатр с большей вероятностью.

Результат оценивания представлен на Рисунке 1.

Задание №6.2.

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

Оценка ковариационной матрицы для нормально распределенных случайных ошибок модели имеет следующий вид:

$$\widehat{\Sigma} = \begin{bmatrix} 1 & -0.4227924 \\ -0.4227924 & 1 \end{bmatrix}$$

То есть коэффициент корреляции между случайными ошибками в двух уравнениях отрицателен. Следовательно, как и было описано в предположениях, может иметь место взаимозаменяемость подписки на онлайн кинотеатр телевидения: наличие подписки снижает вероятность просмотра телевизора, а частый просмотр ТВ уменьшает вероятность наличия подписки на онлайн кинотеатр.

Интерпретируя оценки коэффициентов видно, что все выбранные коэффициенты значимы в том направлении, в котором предполагалось (кроме общей независимой переменной). То есть доход и количество просмотренных сериалов положительно влияют на вероятность наличия подписки, а для вероятности частого просмотра телевидения видно, что чем старше индивид, тем вероятнее он чаще смотрит телевидение. Для того чтобы оценить влияние доля времени в интернете нужно оценивать предельные эффекты, что будет сделано далее.

Рис. 1: Результат оценивания системы уравнений в R

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

EQUATION 1
Link function for mu.1: probit
Formula: sub ~ series + internet + income

Parametric coefficients:
             Estimate Std. Error z value Pr(>|z|)
(Intercept) -1.365e+00  5.638e-02 -24.209   <2e-16 ***
series       7.539e-02  6.389e-03  11.800   <2e-16 ***
internet     9.105e-01  7.161e-02  12.715   <2e-16 ***
income       6.984e-06  7.977e-07   8.756   <2e-16 ***
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1


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

Parametric coefficients:
             Estimate Std. Error z value Pr(>|z|)
(Intercept)  0.0402535  0.0583239    0.69     0.49
age          0.0151176  0.0008115   18.63   <2e-16 ***
internet    -1.2586122  0.0742701  -16.95   <2e-16 ***
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1


n = 5000  theta = -0.423(-0.47,-0.374)  tau = -0.278(-0.312,-0.244)
total edf = 8
```

Задание №6.3.

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

Нулевая гипотеза будет заключаться в том, что корреляция равна 0.

$$\begin{cases} H_0 : \rho = 0 \\ H_a : \rho \neq 0 \end{cases}$$

Статистика теста $LR = 2(\ell_{UR} - \ell_R)$ распределена как хи-квадрат с 1 степенью свободы (одно ограничение на параметр модели).

Проведя тест найденное p-value равно 2.146829e-52, то есть нулевая гипотеза отвергается, а значит, уравнения необходимо оценивать совместно.

Задание №6.4.

Для индивида с произвольными характеристиками оцените:

1. *Вероятность подписки*
2. *Вероятность того, что индивид смотрит телевизор по крайней мере раз в неделю*
3. *Вероятность того, что индивид и имеет подписку, и смотрит телевизор не реже раза в неделю*

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

Возьмем индивида с такими характеристиками, представленные в Таблице 9:

Таблица 9: Характеристики рассматриваемого индивида в системах уравнений

Переменная	Значение
$series_{ind}$	4
$internet_{ind}$	0.7
$income_{ind}$	50000
age_{ind}	21

Посчитаем вероятности для каждого пункта отдельно:

1. $\widehat{P}(sub_{ind} = 1) = \mathcal{F}(\hat{\beta}_0 + \hat{\beta}_1 \times series_{ind} + \hat{\beta}_2 \times internet_{ind} + \hat{\beta}_3 \times income_{ind}) = 0.4693562$

   То есть у индивида вероятность наличия подписки без учета второго уравнения и частоты просмотра ТВ равна примерно 46%.

2. $\widehat{P}(TV_{ind} = 1) = \mathcal{F}(\hat{\gamma}_0 + \hat{\gamma}_1 internet_{ind} + \hat{\gamma}_2 age_{ind}) = 0.3003812$

   То есть у индивида вероятность частого просмотра телевидения без учета первого уравнения и наличия подписки на онлайн кинотеатр равна примерно 30%.

3. $\widehat{P}(sub_{ind} = 1,\, TV_{ind} = 1) = \mathcal{F}(\widehat{sub^*_{ind}}, \widehat{TV^*_{ind}}; , \hat{\rho}) = 0.04156213$

   Где $\widehat{sub^*_{ind}}, \widehat{TV_{ind}}^*$ - оцененные вероятности для индивида в двух уравнениях, а $\mathcal{F}$- двумерное нормальное распределение.

   То есть у индивида вероятность частого просмотра телевидения и наличия подписки на онлайн кинотеатр равна примерно 4%.

4. $\widehat{P}(sub_{ind} = 1 \,|\, TV_{ind} = 0) = \dfrac{\widehat{P}(sub_{ind} = 1, TV_{ind} = 0)}{\widehat{P}(TV_{ind} = 0)} = \dfrac{\widehat{P}(sub_{ind} = 1, TV_{ind} = 0)}{1 - \widehat{P}(TV_{ind} = 1)} = 0.369943$

   То есть у индивида вероятность наличия подписки на онлайн кинотеатра при условии, что он не часто смотрит телевидение равна примерно 37%.

Часть 7. Сравнение моделей

Задание №7.1.

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

Ниже представлена Таблица 10 для сравнения прогностических сил (*accuracy*)моделей.

Таблица 10: Доли верных предсказаний пробит, линейной-вероятностной, наивной модели, логит модели и модели системы бинарных уравнений

Модель	Доля верных предсказаний
Пробит модель	0.6760
Наивная модель	0.6446
Линейно-вероятностная модель	0.6774
Логит модель	0.6754
Система бинарных уравнений	0.6466

Можно сказать, что предсказательная сила пробит и логит моделей лучше наивной модели, но хуже чем у линейной-вероятностной модели. А модель с системой бинарных уравнений и вовсе не сильно лучше наивной модели, что может быть вызвано другой функциональной формой и не учетом важных факторов (как возраст в квадрате).

Таким образом, по предсказательной силе наилучшая модель - линейно-вероятностная модель, хотя как уже было выше отмечено, она трудноинтерпретируема и имеет ряд недостатков.

Задание №7.2.

Выберите лучшую из оцененных вами моделей руководствуясь информационными критериями.

Ниже представлена Таблица 11 для сравнения моделей по информационным критерям AIC, BIC, которые считаются по следующим формулам:

$$\text{AIC} = 2k - 2\ln(\hat{L})$$
$$\text{BIC} = k\ln(n) - 2\ln(\hat{L})$$

Как видно из Таблицы 11, логит и пробит модели лучшие модели с наименьшими показателями по AIC, BIC и практически одинаковы с небольшим перевесом у логит модели. При этом линейно-вероятностная модель хоть и хуже этих моделей (показатели критериев выше), но модель с системой бинарных уравнений намного хуже. Это вызвано тем, что она оценивает сразу два уравнения, поэтому для сравнения нужно и по другим моделям оценить 2 уравнения.

Таблица 11: Общие информационные критерии AIC, BIC для пробит, линейной-вероятностной, наивной модели, логит модели и модели системы бинарных уравнений

Модель	AIC	BIC
Пробит модель	6057.949	6103.570
Линейно-вероятностная модель	6365.681	6417.819
Логит модель	6057.887	6103.507
Система бинарных уравнений	11954.374	12006.511

В таблице 12 представлены результаты после оценивания 2 уравнений каждой из модели и сложением AIC и BIC по двум уравнениям для всех моделей, кроме системы.

Таблица 12: Информационные критерии AIC, BIC для пробит, линейной-вероятностной, наивной модели, логит модели и модели системы бинарных уравнений для 2 выбранных уравнений

Модель	AIC	BIC
Пробит модель	12238.62	12284.24
Линейно-вероятностная модель	12868.13	12926.79
Логит модель	12242.08	12287.70
Система бинарных уравнений	11954.374	12006.511

Можно отметить, что в этом случае логит и пробит модель также близки по информационным критериям с небольшим улучшением у пробит модели в этот раз. Они также лучше линейно-вероятностной, хотя были хуже ее по прогностической силе. Однако модель с системой бинарных уравнений является лучше других моделей (критерии AIC, BIC меньше всего), что объясняется тем, что уравнения необходимо оценивать вместе (даже специально проверяли гипотезу на совместимость оценивания уравнений).

То есть лучшая модель по информационным критериям - модель с системой бинарных уравнений.

Часть 8. Модель бинарного выбора со случайными ошибками, имеющими распределение Стьюдента

Число степеней свободы в данном задании будет равно $df = 13$.

Задание №8.1.***

Используя воображение придумайте и кратко опишите экономическую задачу, для решения которой необходимо применить модель бинарного выбора. Например, можно рассмотреть влияние различных факторов на вероятность дефолта банка. Укажите зависимую переменную и по крайней мере три независимых, а также кратко опишите предполагаемый механизм влияния независимых переменных на зависимую.

Будем рассматривать задачу по определению факторов, влияющих на частое употребление кофе. Это достаточно важная задача особенно для компаний-производителей и самих продавцов кофе, ведь им важно узнать останется ли с ними человек и продолжит часто покупать и пить кофе.

То есть в качестве зависимой переменной будет выступать $coffee$ - дамми переменная на то, что человек пил кофе за последнюю неделю.

В качестве независимых переменных рассмотрим такие факторы, влияющие на употребление кофе:

- $income$ - Доход индивида

  Он влияет на факт употребления кофе, так как кофе является не таким дешевым продуктом и многие просто могут себе не позволить покупать его. Поэтому ожидается, что больше и чаще пьют кофе более люди с более высоким заработком. Учитывая лог-нормальность распределения, то в моделях доход будет рассматривается с использованием логарифмирования.

  Гипотеза: доход положительно влияет на частоту употребления кофе.

- age - Возраст индивида

  Явно влияет на факт употребления кофе, так как чаще пьют кофе те, кто хочет не уснуть и к таким больше относятся молодые люди (студенты) и зрелые (рабочие люди). А более пожилые люди заинтересованы заботой своим здоровьем и будут реже пить кофе.

  Гипотеза: возраст отрицательно влияет на частоту употребления кофе.

- $cigarette$ - Дамми переменная для курящих людей

  Выбрана эта дамми, так как если человек имеет одну зависимость в форме никотина, то скорей всего, будет склонен и к другим зависимостям.

  Гипотеза: факт курения положительно влияет на частоту употребления кофе.

Задание №8.2.***

Симулируйте процесс генерации данных, соответствующий вашей задаче в логике бинарной модели со случайными ошибками, имеющими распределение Стьюдента (по аналогии с тем, как это делалось для обычной пробит модели для дефолта на семинаре). В тексте работы этот пункт отражать не нужно, достаточно реализовать его в коде. Все дальнейшие пункты выполняются на симулированных данных из выборки объемом 5000 наблюдений. Перед началом симуляций необходимо указать set.seed(123).

Модель и истинные значения коэффициентов предположим такими (всего будет 60% любителей кофе в выборке):

$$P(coffee = 1) = 1 + 0.5 \times log(income) - 0.09 \times age + 0.2 \times cigarette$$

Задание №8.3.***

Оцените параметры вашей модели с использованием бинарной модели со случайными ошибками, имеющими распределение Стьюдента с 13 степенями свободы. Результат представьте в форме таблицы, содержащей оценки коэффициентов и p-value тестов на значимость.

Полученные результаты представлены в Таблице 13. Как легко заметить, полученные коэффициенты достаточно близки к истинным значениям и значимы, а также подтверждают вынесенные выше гипотезы.

Таблица 13: Оценки бинарной модели со случайными ошибками, имеющими распределение Стьюдента с 13 степенями свободы

Переменная	$\hat{\beta}$	p-value
Constant	0.43759880	3.249177e-01
log(income)	0.60110278***	0.000000e+00
age	-0.09669452***	1.585945e-279
cigarette	0.17861389***	4.985862e-03