ПРАВИТЕЛЬСТВО РОССИЙСКОЙ ФЕДЕРАЦИИ

Федеральное государственное автономное образовательное учреждение высшего образования

Национальный исследовательский университет «Высшая школа экономики»

Факультет экономических наук
Образовательная программа «Экономика»

ДОМАШНЕЕ ЗАДАНИЕ 1

«Прикладная микроэконометрика»

Студент группы БЭК165
Зехов Матвей Сергеевич

Преподаватель:
Потанин Богдан Станиславович

Москва 2019

Содержание

1 Часть 1. Теория и гипотезы.

☼ 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.). Приведите по крайней мере две ссылки на научные работы (желательно из Q1 журналов), изучавшие влияние различных факторов на вероятность брака. Кратко опишите методологию и основные результаты этих исследований, а также как ваш выбор переменных и сформулированные гипотезы соотносятся с данными результатами.

Постарайтесь подбирать независимые переменные таким образом, чтобы избежать проблемы обратной причинности. То есть ваши независимые переменные должны, в теоретическом смысле, влиять на зависимую, а не наоборот. Например, уместно предположить, что возраст влияет на вероятность брака. В то же время, скорее всего, не количество детей влияет на вероятность брака, а, вполне возможно, наоборот. Поэтому возраст можно включить в число независимых переменных, а число детей — не желательно

☼ 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.

☼ Определитесь с тем, будете ли вы предсказывать вероятность брака для мужчин и женщин в рамках единой модели, либо остановитесь лишь на одном из полов. Обоснуйте свой выбор теоретически.

В этой работе будет рассматриваться вероятность вступления в брак для совершеннолетних мужчин. Для этого были выбраны следующие переменные: возраст в качестве непрерывной переменной и наличие работы в качестве дамми-переменной, а также возраст в квадрате и произведение двух базовых переменных. Модель для мужчин будет оцениваться отдельно от женщин, потому что поведение мужчин и женщин не только отличается возрастом заключения первого брака (Goodwin, P., McGill, B., & Chandra, A. (2009). Who marries and when?; age at first marriage in the United States, 2002.), но и направление эффектов может различаться по знаку, так как мужчины и женщины воспринимают брак по-разному.

С возрастом вероятность быть в браке для мужчин повышается, о чём говорят многие исследования. Для проверки гипотезы о наличии нелинейной зависимости от возраста была введена переменная квадрата возраста. Можно предположить, что вероятность вступления в брак максимальна для мужчин среднего возраста, о чём бы говорил значимость этого коэффициента. Трудоустройство мужчины считается одним из самых сильных показателей при прогнозировании вероятность его семейного статуса: работающие мужчины с большей вероятностью будут женаты. Стоит отметить, что в качестве предиктора стоит использовать именно факт наличия работы, а не величину дохода, потому что существует неоднозначный эффект супружеской премии. Направление зависимости до конца не изучено. Существуют исследования, согласно которым факт замужества может влиять на величину труда (Аистов, А. В., & Коваленко, Н. В. Супружеская премия//XIV Апрельская международная научная конференция по проблемам развития экономики и общества: в 4 кн./отв. ред. ЕГ Ясин, 661-671.), так и наоборот (Chun, H., & Lee, I. (2001). Why do married men earn more: Productivity or marriage selection?. Economic Inquiry, 39(2), 307-319.). Переменная произведения трудоустройства и возраста позволит изучить совместное

влияние двух этих факторов на вероятность вступления в брак. Можно предположить, что для более молодых мужчин наличие работы будет фактором, снижающим вероятность нахождения в браке из-за того, что они либо не ходят в университет, либо совмещают работу и учёбу, а потому у них остаётся меньше времени для формирования семьи. Для более старших мужчин, напротив, наличие работы будет важным показателем, так как это позволит им обеспечивать семью и, возможно, сделает их более привлекательными для заключения брака.

В статье «Sex Differences in the Entry into Marriage» (Goldscheider, F. K., & Waite, L. J. (1986). Sex differences in the entry into marriage. American journal of sociology, 92(1), 91-109.) авторы оценивали логистическую модель на выборке 18-29 лет для мужчин, которую они разбили на несколько кластеров для предсказания вероятности «ранней» и «поздней» женитьбы. В качестве объясняющих переменных авторы выбрали расу, образование, трудоустройство, военную службу, образование родителей. В результате они обнаружили, что влияние трудоустройства на вероятность заключения брака положительно для всех возрастов, но сильнее для более старших индивидов. В другой статье «Employment and marriage among inner-city fathers» (Testa, M., Astone, N. M., Krogh, M., & Neckerman, K. M. (1989). Employment and marriage among inner-city fathers. The Annals of the American Academy of Political and Social Science, 501(1), 79-91.) авторы анализировали, будет ли мужчина жениться на женщине после зачатия с ней первого ребёнка. Для этого авторы строят логистическую регрессию для предсказания вероятности заключения брака в период между зачатием и рождением ребёнка на основе расы, трудоустройства и образования мужчины и других параметров. В результате было замечено, что у трудоустроенного отца в два раза выше вероятность легитимизировать брак до рождения ребёнка.

2 Часть 2. Обработка данных

2.1

☼Приведите ваши данные в порядок, удалив пропуски и очистив используемые переменные от неточностей (осуществляется в коде, в работе это описывать не обязательно).

2.2

☼Рассмотрите доли холостых людей и состоящих в браке. Укажите в таблице описательные статистики для ваших независимых переменных: отдельно для холостых и состоящих в браке. Сделайте выводы о различии в распределении независимых переменных среди индивидов в зависимости от семейного статуса

Ответ. Доля состоящих в браке людей составляет 61% от выборки. В таблицах 1 и 2 приведены описательные статистики для женатых и холостых мужчин.

На основе приведённых статистик можно сделать следующие выводы:

- ☼ Средний и медианный возраст женатых мужчин выше, чем неженатых. В общем, ожидаемый результат. Он согласуется с озвученным выше предположением о том, что для мужчин с возрастом растёт вероятность вступления в брак.
- ☼ Доля занятых среди женатых выше, чем среди неженатых. Скорее всего это связано с необходимостью для мужчины обеспечивать семью и поддерживать достаточный уровень жизни. Женатым сложнее оставить на некоторое время работу, даже при наличии возможности, так как от них зависят и члены семьи.

age	work	alc	male	marriage	workxage	educ_high	educ_mid	educ_other
Min. :18.00	Min. :0.0000	Min. :0.0000	Min. :1	Min. :1	Min. : 0.00	Min. :0.0000	Min. :0.0000	Min. :0.0000
1st Qu.:38.00	1st Qu.:0.0000	1st Qu.:1.0000	1st Qu.:1	1st Qu.:1	1st Qu.: 0.00	1st Qu.:0.0000	1st Qu.:0.0000	1st Qu.:0.0000
Median :51.00	Median :1.0000	Median :1.0000	Median :1	Median :1	Median :34.00	Median :0.0000	Median :0.0000	Median :0.0000
Mean :50.64	Mean :0.6439	Mean :0.7522	Mean :1	Mean :1	Mean :28.29	Mean :0.2716	Mean :0.3791	Mean :0.3493
3rd Qu.:62.00	3rd Qu.:1.0000	3rd Qu.:1.0000	3rd Qu.:1	3rd Qu.:1	3rd Qu.:47.00	3rd Qu.:1.0000	3rd Qu.:1.0000	3rd Qu.:1.0000
Max. :89.00	Max. :1.0000	Max. :1.0000	Max. :1	Max. :1	Max. :77.00	Max. :1.0000	Max. :1.0000	Max. :1.0000

Таблица 1: Описательные статистики для женатых

age	work	alc	male	marriage	workxage	educ_high	educ_mid	educ_other
Min. :18.00	Min. :0.0000	Min. :0.0000	Min. :1	Min. :0	Min. : 0.00	Min. :0.0000	Min. :0.0000	Min. :0.0000
1st Qu.:25.00	1st Qu.:0.0000	1st Qu.:0.0000	1st Qu.:1	1st Qu.:0	1st Qu.: 0.00	1st Qu.:0.0000	1st Qu.:0.0000	1st Qu.:0.0000
Median :33.00	Median :1.0000	Median :1.0000	Median :1	Median :0	Median :22.00	Median :0.0000	Median :0.0000	Median :0.0000
Mean :39.12	Mean :0.5487	Mean :0.6881	Mean :1	Mean :0	Mean :19.47	Mean :0.1762	Mean :0.3663	Mean :0.4575
3rd Qu.:51.00	3rd Qu.:1.0000	3rd Qu.:1.0000	3rd Qu.:1	3rd Qu.:0	3rd Qu.:33.00	3rd Qu.:0.0000	3rd Qu.:1.0000	3rd Qu.:1.0000
Max. :92.00	Max. :1.0000	Max. :1.0000	Max. :1	Max. :0	Max. :77.00	Max. :1.0000	Max. :1.0000	Max. :1.0000

Таблица 2: Описательные статистики для холостых

☼ Доля людей, употребляющих алкоголь, выше среди женатых, хотя это не слишком сильно выражено. Возможно, эта разница может быть связана с семейными проблемами

☼ Среди женатых выше доля людей с высшим образованием. Это может быть объяснено тем, что наличие ступени обучения повышает вероятность брака в принципе.

2.3

☼Найдите характеристики среднего и медианного индивидов. Результат представьте в форме таблицы и опишите его словами. Сделайте выводы о различии в характеристиках между средним и медианным индивидами.

Ответ. В Таблице 3 представлены характеристики среднего и медианного индивидов. Как можно видеть, если округлить средние значения бинарных переменных, средний и медианный индивид почти не различаются. Возраст среднего индивида чуть выше, чем у медианного. У обоих есть работа и оба употребляют алкоголь. Также оба женаты.

	age	work	alc	male	marriage	workxage	educ_high	educ_mid	educ_other
Mean	46.17	1	1	1	1	24.86	0	0	0
Median	45.00	1	1	1	1	28.00	0	0	0

Таблица 3: Характеристики среднего и медианного индивидов

2.4

☼Отобразите при помощи двух гистограмм различия в распределении одной из ваших независимых переменных в зависимости от семейного статуса индивида. Сделайте выводы о различиях в распределении, а также приведите возможное теоретическое обоснование данных различий.

Ответ. Для визуализации рассмотрим гистограммы возраста в зависимости от семейного положения индивида. Их можно увидеть на Рис. 1. Как легко заметить, среднее распределения неженатых мужчин действительно смещено влево, но имеет при этом тяжёлый хвост. Это говорит о том, что неженатыми в основном является молодёжь. В целом, это логичный результат, что молодёжь не слишком торопится вступать в брак. Тем не менее,

тяжёлый хвост говорит о том, что довольно большой процент мужчин вообще не может в ступить в брак по каким-либо причинам, либо разведены.

Гистограмма распределения женатых мужчин говорит о том, что распределение довольно симметрично. Не слишком много мужчин рано вступает в брак и, очевидно, не слишком много доживает до преклонных лет. Хвосты распределения тоже тяжёлые, но не так сильно, как для неженатых.

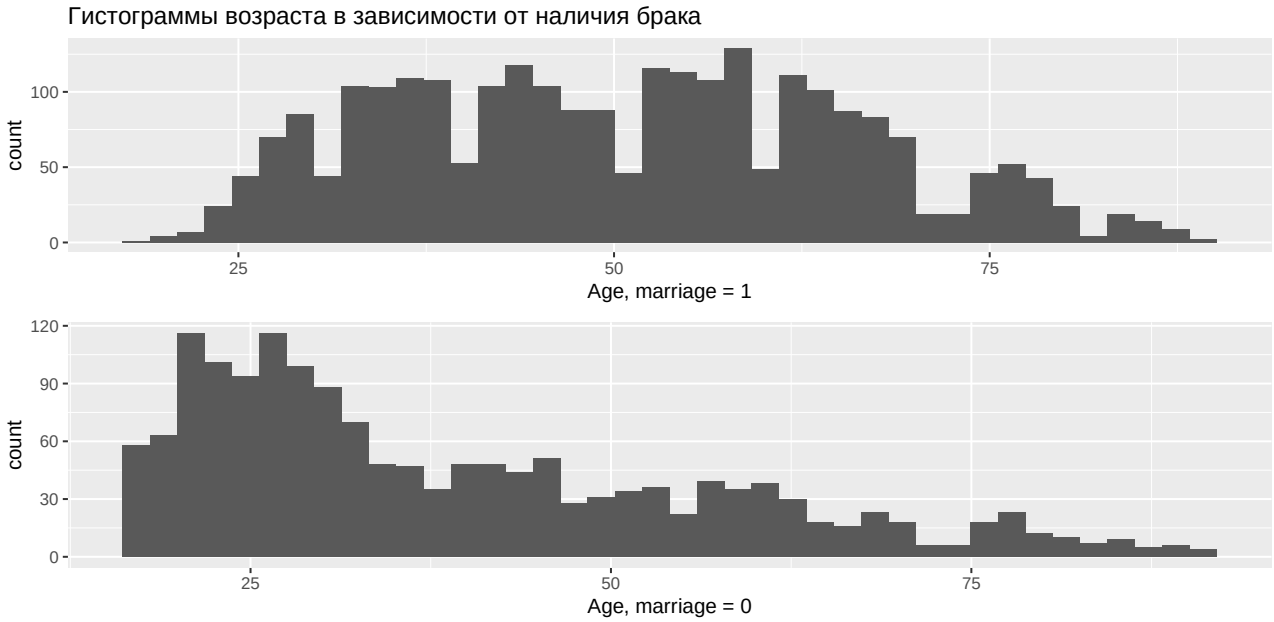

Рис. 1: Гистограммы

3

3.1

☼Оцените линейно-вероятностную модель, предварительно записав регрессионное уравнение. Результат представьте в форме таблицы (можно, например, использовать выдачу из stata, R или python).

Ответ. Регрессионное уравнение линейно-вероятностной модели:

$$marriage_i = \beta_0 + \beta_1 age_i + \beta_2 age_i^2 + \beta_3 work_i + \beta_4 workxage_i + \varepsilon_i$$

В Таблице 4 представлены результаты оценки этой модели.

	Estimate	Std. Error	t value	Pr(>|t|)	Metrics	Value
(Intercept)	-0.9034	0.0609	-14.83	0.0000	R^2	0.21
age	0.0514	0.0026	19.86	0.0000	Adj. R^2	0.21
I(age^2)	-0.0004	0.0000	-15.68	0.0000	Num. obs.	4122
work	0.3648	0.0442	8.26	0.0000	RMSE	0.43
workxage	-0.0047	0.0010	-4.90	0.0000	F-statistic(pval)	274.2(0.0000)

Таблица 4: Результаты оценки линейно-вероятностной модели

3.2

☼Скорректируйте оценку ковариационной матрицы оценок регрессионных коэффициентов при помощи бутстрапа. Объясните причины, по которым необходимо осуществлять данную корректировку и к каким негативным последствиям может привести её отсутствие. Результат представьте в форме таблицы (можно, например, использовать выдачу из stata, R, python).

Ответ. Так как предположение о нормальности ошибок может не соблюдаться, то метод максимального правдободобия может давать несостоятельные и смещённые оценки коэффициентов и ковариационной матрицы, так как правдоподобие будет записано ошибочно. Бутстрап помогает преодолеть эту проблему и получить состоятельные оценки без предположения о нормальности ошибок. Результаты можно видеть в Таблице 5.

	Бутстрапированные s.e.	Старые s.e.
(Intercept)	0.0430	0.0609
age	0.0022	0.0026
I(age^2)	0.0000	0.0000
work	0.0370	0.0442
workxage	0.0008	0.0010

Таблица 5: Бутстрапированные и старые стандартные ошибки

3.3

☼Протестируйте гипотезы о значимости коэффициентов при помощи перцентильных бутстрапированных доверительных интервалов и сравните полученный вами результат с тем, что был получен при изначальном оценивании (без корректировки). Результат представьте в форме таблицы (можно, например, использовать выдачу из stata, R, python). Поясните, как полученные результаты соотносятся с высказанными вами ранее предположениями.

Ответ. Для тестирования гипотез построим бутстрапированные доверительные интервалы. Результаты можно видеть в Таблице 6. Последний столбец показывает, отвергается ли гипотеза о значимости коэффициента. Так как ноль не входит ни в один интервал, гипотеза всюду отвергается. По сравнению с обычной оценкой ковариационной матрицы, ничего не изменилось.

	Left	Right	Coef	Rejected
Intercept	-0.9873	-0.8183	-0.9034	TRUE
Age	0.0465	0.0558	0.0514	TRUE
Age^2	-0.0004	-0.0003	-0.0004	TRUE
Work	0.2921	0.4616	0.3648	TRUE
Workage	-0.0068	-0.0034	-0.0047	TRUE

Таблица 6: Бутстрапированные доверительные интервалы

3.4

☼Запишите формулу, по которой можно рассчитать предельные эффекты в линейно-вероятностной модели для:

А) Непрерывной переменной, входящей в регрессионное уравнение линейно.

Б) Дамми-переменной.

В) Непрерывной переменной, входящей в регрессионное уравнение кубически.

Г) Непрерывной переменной, имеющей взаимодействие с другой переменной.

Ответ. Для каждого случая будем предварительно записывать рассматриваемую модель.

☼ Предельный эффект по непрерывной переменной

$$\hat{y}_i = \hat{\beta}_0 + \hat{\beta}_1 x_{1i}$$

$$\frac{d\hat{y}_i}{dx_{1i}} = \hat{\beta}_1$$

☼ Предельный эффект по дамми-переменной

$$\hat{y}_i = \hat{\beta}_0 + \hat{\beta}_1 d_{1i}$$

$$\frac{d\hat{y}_i}{dd_{1i}} = \hat{y}_i(d_{1i} = 1) - \hat{y}_i(d_{1i} = 0) = \hat{\beta}_0 + \hat{\beta}_1 - \hat{\beta}_0$$

☼ Предельный эффект по непрерывной переменной

$$\hat{y} = \hat{\beta}_0 + \hat{\beta}_1 x_{1i}^3$$

$$\frac{d\hat{y}_i}{dx_{1i}} = 3\hat{\beta}_1 x_{1i}^2$$

☼ Предельный эффект для непрерывной переменной, имеющей взаимодействие с другой переменной.

$$\hat{y}_i = \hat{\beta}_0 + \hat{\beta}_1 x_{1i} x_{2i} + \hat{\beta}_2 x_{2i}$$

$$\frac{d\hat{y}_i}{dx_{1i}} = \hat{\beta}_1 x_{2i}$$

3.5

☼Проинтерпретируйте полученные значения оценок предельных эффектов для каждой независимой переменной. Значения оценок предельных эффектов должны быть представлены в форме таблицы с указанием соответствующих значений оценок стандартных отклонений (стандартных ошибок). Исключение составляют лишь независимые переменные, входящие в регрессионное уравнений нелинейно: для них стандартные ошибки считать не обязательно.

Ответ. Неясно, что тут требуется проинтерпретировать, так как все переменные входят нелинейно и все индивиды имеют свои предельные эффекты. Для того, чтобы тут что-то было, представлю верх таблицы с предельными эффектами. Его можно найти в Таблице 7.

ID	Age	Work
5	0.37	0.09
9	0.37	0.08
14	0.37	0.08
17	0.38	0.15
19	0.37	0.11
23	0.00	0.09

Таблица 7: Предельные эффекты для нескольких индивидов

3.6

☼Посчитайте средний предельный эффект, предельный эффект для среднего индивида, предельный эффект для медианного индивида и предельный эффект для индивида с вашими характеристиками по каждой из независимых переменных, представив результат в форме таблицы.

Ответ. Как можно видеть из Таблицы 8, предельные эффекты среднего и медианного индивидов очень близки. Этот результат логичен, так как модель линейна, а характеристики среднего и медианного индивидов тоже близки. Как легко видеть, оба эффекта положительны. Это говорит о том, что при увеличении возраста среднего индивида вероятность его вступления в брак повышается. Аналогично и с занятостью. При наличии работы вероятность так же повышается. Так, при росте возраста среднего индивида на 1, вероятность его вступления в брак повышается на 0.38, а при переходе из состояния "незанятый"в состояние "занятый"эта вероятность возрастает почти на 0.15. Почти аналогичны результаты для медианного индивида.

Что же касается меня, то при увеличении моего возраста на год, вероятность брака возрастает всего на 3%. Однако если я найду работу, то она возрастает почти на 27%.

	Age	Work
Mean ME	0.2361	0.1496
ME of mean person	0.3795	0.1496
ME of median person	0.3804	0.1551
ME of Me	0.0347	0.2669

Таблица 8: Предельные эффекты: средний, для среднего и медианного индивидов

3.7

☼Посчитайте стандартные ошибки оценок предельных эффектов независимых переменных, входящих в регрессионное уравнений нелинейно (не считая переменные взаимодействия), для индивида с вашими характеристиками, предварительно пояснив методологию расчета.

Закроем глаза на то, что возраст также включён в переменную взаимодействия, чтобы соблюсти условия пункта. Для получения стандартных ошибок, выпишем формулу дисперсии предельного эффекта для возраста.

$$\mathrm{Var}\left(\frac{d\hat{y}}{dage}\right) = \mathrm{Var}(\hat{\beta}_0 + 2\hat{\beta}_1 x_i) = \mathrm{Var}(\hat{\beta}_0) + 4x_i^2\,\mathrm{Var}(\hat{\beta}_1) + 4x_i cov(\hat{\beta}_0, \hat{\beta}_1) =$$

Так как истинную дисперсию мы не знаем и никогда не узнаем, то оценим её, взяв оценки дисперсий и ковариации из оценённой ковариационной матрицы. Для индивида с моими характеристиками, представленными в Таблице 10, это значение равно 0.001562

3.8

☼Посчитайте долю верных предсказаний и сопоставьте её с результатом наивного прогноза. Сделайте вывод о предсказательной силе полученной модели.

Ответ. Доля верных предсказаний линейной и наивной моделей представлена в Таблице 9. Как легко видеть, линейная модель работает существенно лучше наивной.

	Linear	Naive
Accuracy	0.7285	0.6118

Таблица 9: Доля верных прогнозов моделей

3.9

Недостатки линейной модели:

☼ Гетероскедастичность по построению модели

☼ Прогнозы модели могут лежать вне отрезка [0,1], что является алогичным и сложно интерпретируемым.

Доля прогнозов, больше 1 или меньше 0: 0.0354

Касательно R^2, на примере парной регрессии в обычной задаче регрессии доказывалось, что R^2 равен выборочному коэффициенту корреляции зависимой и независимой переменных. В случае, когда зависимая бинарная, а независимая - непрерывная или бинарная, некорректно считать обычную корреляцию и интерпретировать её. Необходим подсчёт тетрахорической корреляции. Следовательно, при регрессии на бинарную переменную R^2 интерпретировать нельзя. Так как $F-$статистика выражается через R^2, она так же неинтерпретируема. Кроме того, ошибки распределены не нормально, следовательн F-статистика бессмысленна.Нужно использовать альтернативные показатели (MacFadden, etc.)

3.10

☼Оцените вероятность брака для индивида с вашими характеристиками. Если вы оцениваете модель по представителям противоположного пола, то оцените соответствующую вероятность для индивида с вашими характеристиками, но противоположного пола

Ответ. Прогноз для индивида с моими характеристиками (представлены в Таблице 10) равен 0.0009

	Intercept	Age	Age^2	Work	Workxage
Me	1	21	441	0	0

Таблица 10: Мои характеристики

4

4.1

Пусть x_i - вектор характеристик i-го индивида ($Intercept\ Age\ Age^2\ Work\ Workxage$), а β - вектор коэффициентов. y_i - бинарная переменная, семейное положение индивида. Функция правдоподобия:

$$L = \prod_i \left[F\left(x_i'\beta\right)\right]^{y_i} \left[1 - F\left(x_t'\beta\right)\right]^{1-y_i}$$

Максимизируемый логарифм функции правдоподобия

$$l = \ln L = \sum_i \left[y_i \ln F\left(x_i'\beta\right) + (1 - y_i) \ln\left(1 - F\left(x_i'\beta\right)\right)\right]$$

$$F(u) = \frac{1}{\sqrt{2\pi}} \int_{-\infty}^{u} e^{-\frac{z^2}{2}} dz$$

Результаты оценки модели представлены в Таблице 11.

	Estimate	Std. Error	z value	Pr(>|z|)
(Intercept)	-5.0420	0.2778	-18.15	0.0000
age	0.1783	0.0102	17.44	0.0000
I(age^2)	-0.0014	0.0001	-15.04	0.0000
work	1.6845	0.1834	9.19	0.0000
workxage	-0.0249	0.0037	-6.65	0.0000

Таблица 11: Результаты оценки probit-модели

4.2

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

Ответ. Все коэффициенты значимы. Проинтерпретируем знаки коэффициентов. Возраст и наличие работы имеют положительный эффект, как и предполагалось в первой части. Произведение возраста на работу имеет слабый отрицательный коэффициент, что говорит о том, что работающим мужчинам в возрасте может быть сложнее вступить в брак. Коэффициент перед квадратом сложно интерпретировать, скорее он имеет роль в предельных эффектах.

4.3

Ответ. Запишите формулу, по которой можно рассчитать предельные эффекты в пробит модели для:

А) Непрерывной переменной, входящей в регрессионное уравнение линейно.

Б) Дамми-переменной.

В) Непрерывной переменной, входящей в регрессионное уравнение кубически.

Г) Непрерывной переменной, имеющей взаимодействие с другой переменной.

Ответ. F - функция распределения стандартного нормального распределения, f - его функция плотности.

☼ Предельный эффект по непрерывной переменной

$$\hat{y}_i = F(\hat{\beta}_0 + \hat{\beta}_1 x_{1i})$$

$$\frac{d\hat{y}_i}{dx_{1i}} = \hat{\beta}_1 f_(\hat{\beta}_0 + \hat{\beta}_1 x_{1i})$$

☼ Предельный эффект по дамми-переменной

$$\hat{y}_i = F(\hat{\beta}_0 + \hat{\beta}_1 d_{1i})$$

$$\frac{d\hat{y}_i}{dd_{1i}} = F(\hat{y}_i(d_{1i} = 1)) - F(\hat{y}_i(d_{1i} = 0)) = F(\hat{\beta}_0 + \hat{\beta}_1) - F(\hat{\beta}_0)$$

☼ Предельный эффект по непрерывной переменной

$$\hat{y} = F(\hat{\beta}_0 + \hat{\beta}_1 x_{1i}^3)$$

$$\frac{d\hat{y}_i}{dx_{1i}} = 3\hat{\beta}_1 x_{1i}^2 f(\hat{\beta}_0 + \hat{\beta}_1 x_{1i}^3)$$

☼ Предельный эффект для непрерывной переменной, имеющей взаимодействие с другой переменной.

$$\hat{y}_i = F(\hat{\beta}_0 + \hat{\beta}_1 x_{1i} x_{2i} + \hat{\beta}_2 x_{2i})$$

$$\frac{d\hat{y}_i}{dx_{1i}} = \hat{\beta}_1 x_{2i} f(\hat{\beta}_0 + \hat{\beta}_1 x_{1i} x_{2i} + \hat{\beta}_2 x_{2i})$$

4.4

☼Рассчитайте значения оценок предельных эффектов для каждой из независимых переменных для каждого индивида в выборке (результат достаточно представить в коде, в тексте описывать это не обязательно). Для переменной, входящей в регрессионное уравнение нелинейно, постройте гистограмму, отражающую распределение значений оценок её предельного эффекта. Запишите, какие значения оценок предельных эффектов для данной переменной встречаются чаще: положительные или отрицательные.

Ответ. Гистограммы предельных эффектов представлены на Рис. 2. Согласно подсчётам, все предельные эффекты по возрасту положительны. Доля положительных предельных эффектов по занятости составляет 0.879. Соответственно, большинство предельных эффектов по обеим независимым переменным положительны.

4.5

Посчитайте стандартные ошибки оценок предельных эффектов независимых переменных, входящих в регрессионное уравнений нелинейно, для индивида с вашими характеристиками, предварительно пояснив методологию расчета.

Подсчитаем дисперсию предельного эффекта.

$$\mathrm{Var}((\beta_1 + 2\beta_2 age_i + \beta_4 work_i) f(latent\_value_i)) = (\mathrm{Var}(\beta_1) + 2age^2 \mathrm{Var}(\beta_2) + work_i^2 \mathrm{Var}(\beta_4) +$$

$$4age_i \mathrm{Cov}(\beta_1, \beta_2) + 2work_i \mathrm{Cov}(\beta_1, \beta_4) + 4age_i work_i \mathrm{Cov}(\beta_2, \beta_4)) * f(latent\_value_i)^2$$

Подставляя на место истинных дисперсий оценки, получим оценку дисперсии предельных эффектов. Для индивида с моими характеристиками (см. Таблицу 10), корень из этой дисперсии равен 0.0004876624.

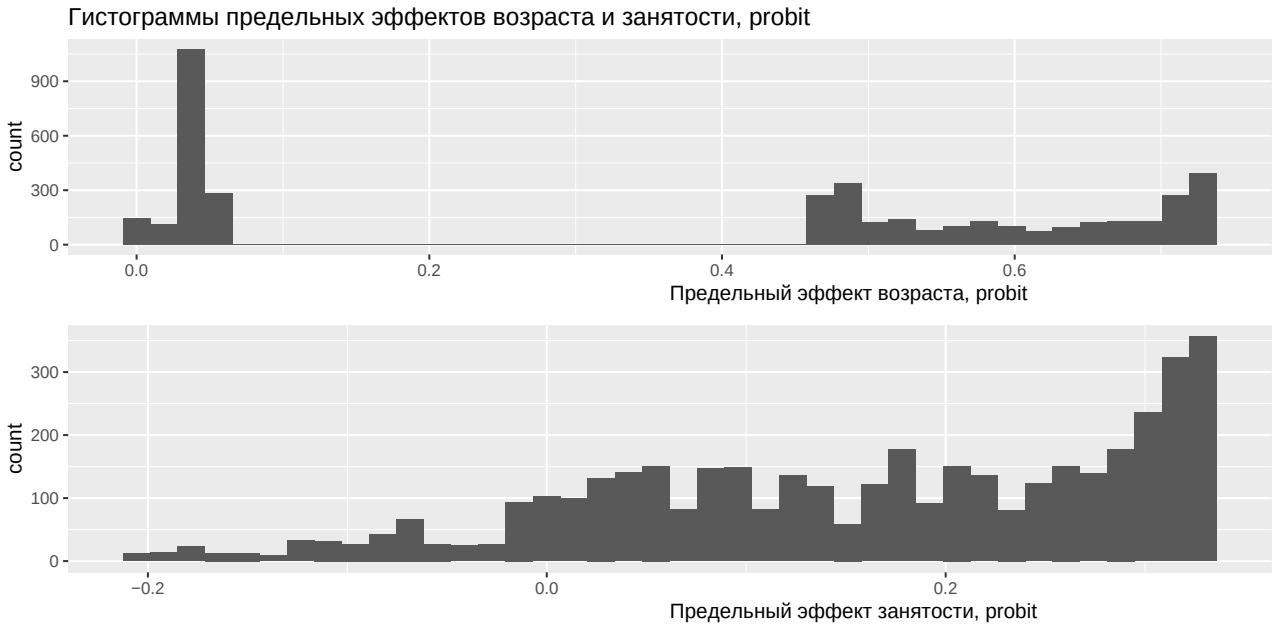

Рис. 2:

4.6

☼Посчитайте значения оценок среднего предельного эффекта, предельного эффекта для среднего индивида, предельного эффекта для медианного индивида и предельного эффекта для индивида с вашими характеристиками для каждой из независимых переменных, представив результат в форме таблицы

Средний предельный эффект, предельные эффекты для среднего и медианного индивидов, а такэе предельные эффекты для меня представлены в Таблице 12

	Age	Work
Mean ME	0.3805	0.1615
ME of mean person	0.4517	0.2084
ME of median person	0.3864	0.1975
ME of Me	0.0079	0.2005

Таблица 12: Предельные эффекты: средний, для среднего и медианного индивидов и меня

4.7

4.8

☼Посчитайте долю верных предсказаний и сопоставьте её с результатом наивного прогноза. Сделайте вывод о предсказательной силе полученной модели

Из Таблицы 13 очевидно, что прогнозная сила probit-модели заметно выше наивной, но ниже линейной.

	Probit	Naive
Accuracy	0.6684	0.6118

Таблица 13: Доля верных прогнозов моделей

4.9

☼Посчитайте значение оценки вероятности брака для индивида с вашими характеристиками. Если вы оцениваете модель по представителям противоположного пола, то оцените соответствующую вероятность для индивида с вашими характеристиками, но противоположного пола.

Ответ. Для индивида с характеристиками, представленными в Таблице 10, прогноз равен 0.0653

5

5.1

☼При помощи LM-теста протестируйте гипотезу о соблюдении допущения о распределении случайных ошибок: при этом необходимо формально записать нулевую и альтернативную гипотезы, статистику теста и её распределение. Запишите, к каким негативным последствиям может привести нарушение данного допущения.

Ответ. Без предположения о нормальности модель, оценённаяя методом максимального правдоподобия, даст смещённые и несостоятельные оценки. Это очевидно следует из того, что правдоподобие оцениваемой модели будет отличаться от правдоподобия истинной модели.

Пусть x_i - вектор характеристик i-го индивида ($Intercept\ Age\ Age^2\ Work\ Workxage$), а β - вектор коэффициентов. y_i - бинарная переменная, семейное положение индивида.

$$\mathbb{P}\{y_i = 1\} = F\left(x_i'\beta + \gamma_1 (x_i;\beta)^2 + \gamma_2 (x_i;\beta)^3\right)$$

$H_0 : \gamma_1 = 0, \gamma_2 = 0$
$H_A : \gamma_1 \neq 0$ or/and $\gamma_2 \neq 0$
Статистика теста:

$$nR^2 \sim \chi^2_2$$

где R^2 из регрессии вектора единиц на вектора первых производных функции правдоподобия каждого индивида по $\beta, \gamma_1, \gamma_2$ (каждый вектор - n-мерный):

$$\ln L_i = \ln F\left(x_i', \beta + r_i (x_i'\beta)^2 + \gamma_i (x_i'\beta)^3\right)^{y_i} \cdot (1 - F(\cdots))^{1-y_i}$$

5.2

☼Предположите, какие переменные могут влиять на дисперсию случайной ошибки. При помощи LR и LM тестов протестируйте гипотезу о гомоскедастичности случайных ошибок: при этом необходимо формально записать нулевую и альтернативную гипотезы, статистику теста и её распределение. Запишите, к каким негативным последствиям может привести нарушение данного допущения. Объясните преимущество LM теста над LR тестом в данном случае.

Ответ. Предположим, что гетероскедастичность в модели порождается переменной *Age*. Например, можно предположить, что в молодом возрасте индивиды более склонны создавать или разрушать семьи, то есть изменять своё семейное положение. Люди старшего возраста интуитивно более инертны в этом плане. Следственно, предполагаем, что дисперсия случайных ошибок будет ниже с ростом возраста. Для визуальной проверки этого предположения построим график квадратов остатков probit-модели, агрегированных (суммированием) по возрастам. Как видно из Рис. 3 , этот ряд убывает с ростом

возраста. Игнорирование факта наличия гетероскедастичности может привести к неэффективности оценок коэффициентов.

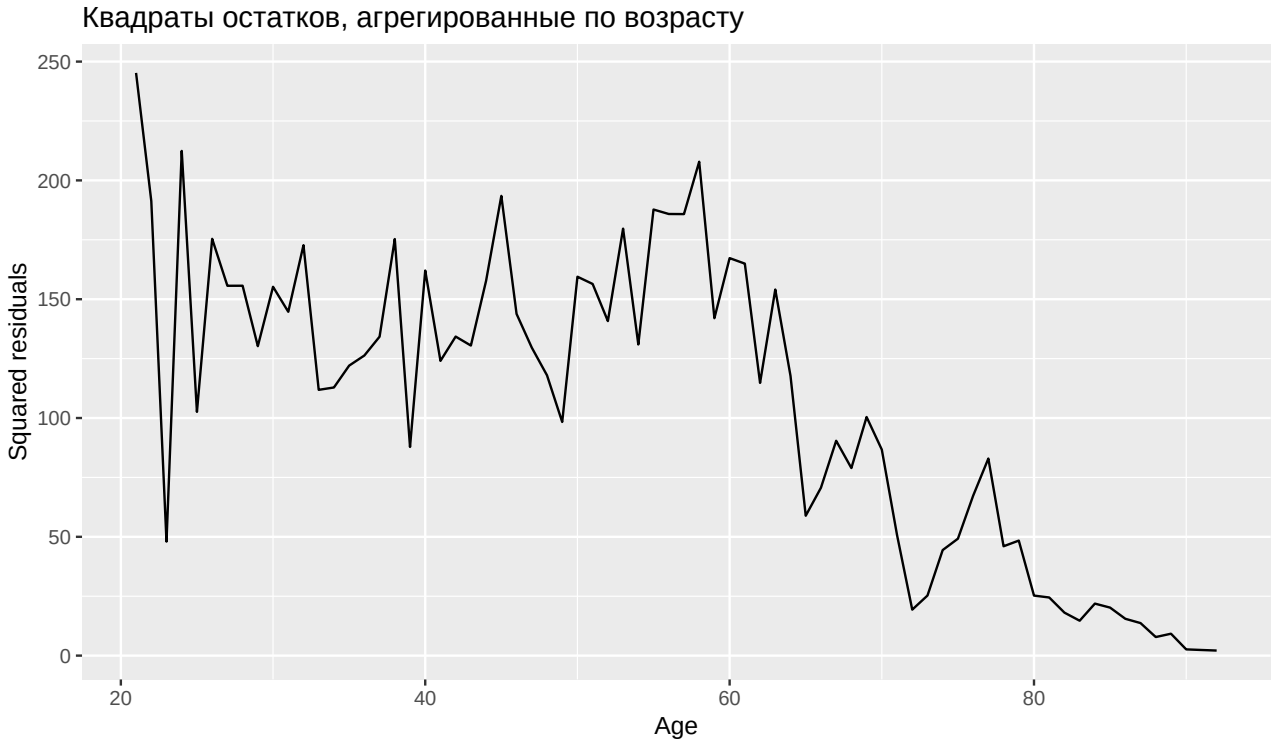

Рис. 3:

5.2.1 LR-тест

Неограниченная модель - Hetprobit

Ограниченная модель - Probit

H_0 - гомоскедастичность $\rightarrow$ модели эквивалентны

H_A - отсутствие гомоскедастичности, необходима Hetprobit-модель.

Тестовая статистика:

$$LR = 2(lnL_{UR} - lnL_R) \sim \chi^2_1$$

Df $= 1$ так как накладывается всего одно линейное ограничение на Hetprobit-модель, а именно, равенство коэффициента перед Age нулю в уравнении дисперсии.

По результатам теста нулевая гипотеза отвергается, так как p-value $= 0$

5.2.2 LM-тест

Пусть x_i - вектор характеристик i-го индивида ($Intercept\ Age\ Age^2\ Work\ Workxage$), z_i - вектор характеристик, подозреваемых в порождении гетероскедастичности, а β - вектор коэффициентов. y_i - бинарная переменная, семейное положение индивида.

$\mathrm{Var}(\varepsilon_i) = h(x_i')k, h(0) = 1', h'(0) \neq 0$

$H_0 : \alpha = 0$

$H_A : \alpha \neq 0$

Статистика теста:

$$nR^2 \sim \chi^2_1$$

где R^2 из регрессии вектора единиц на вектор первых производных правдоподобия по каждому индивиду UR-модели по β, α. Одна степеръ свободы, так как только одна переменная подозревается в порождении гетероскедастичности.

5.3

☼Предпримите действия для борьбы с гетероскедастичностью, оценив учитывающую её модель. Сравните знаки и значимости коэффициентов с теми, что были получены при оценивании модели без учета гетерокедастичности.

Ответ. В Таблицах 14 и 15 приведены результаты оценки Hetprobit- и Probit-моделей. Как можно видеть, стандартные ошибки всех коэффициентов выросли, а значит могли измениться результаты тестирования значимости коэффициентов. Однако, этого не произошло. Все коэффициенты остались значимы на уровне 1%. Только у переменной *workxage* слегка выросло p-value. Теперь, после коррекции гетероскедастичности, оценки стали более эффективными. Знаки коэффициентов не изменились.

	Estimate	Std. Error	z value	Pr(>|z|)
(Intercept)	-8.0593	1.2591	-6.4008	0.0000
age	0.2825	0.0403	7.0155	0.0000
I(age^2)	-0.0020	0.0002	-8.1554	0.0000
work	2.2055	0.4106	5.3709	0.0000
workxage	-0.0262	0.0085	-3.0761	0.0021

Таблица 14: Результат оценивания Hetprobit-модели

	Estimate	Std. Error	z value	Pr(>|z|)
(Intercept)	-5.0420	0.2778	-18.15	0.0000
age	0.1783	0.0102	17.44	0.0000
I(age^2)	-0.0014	0.0001	-15.04	0.0000
work	1.6845	0.1834	9.19	0.0000
workxage	-0.0249	0.0037	-6.65	0.0000

Таблица 15: Результат оценивания Probit-модели

	Hetprobit	Probit	Naive
Accuracy	0.7276	0.6684	0.6118

Таблица 16: Доля верных прогнозов моделей

5.4

☼Сравните долю верных прогнозов пробит модели с гетероскедастичной случайной ошибкой с долей верных прогнозов, полученных при помощи обычной пробит модели (для тех, кто работает в STATA, это задание со звездочкой *).

Ответ. Сравним качество прогнозов Hetprobit- и Probit-моделей. В Таблице 16 приведены доли верных ответов для обеих моделей. Как можно видеть, качество существенно возросло, но оно всё ещё не переигрывает линейную модель.

5.5

5.6

5.7

☼При помощи LR теста проверьте, можно ли оценивать совместную модель для людей с различными уровням образования, либо стоит оценить три различные модели для индивидов с 1) высшим, 2) средним или средним специальным 3) иным уровнем образования. При помощи LR теста проверьте, можно ли оценивать совместную модель для людей с различными уровням образования, либо стоит оценить три различные модели для индивидов с 1) высшим, 2) средним или средним специальным 3) иным уровнем образования.

Ответ.

Нулевая гипотеза: Обе модели верны

Альтернативная гипотеза: Ограниченная модель неверна, необходимо использоать неограниченную.

Нулевая гипотеза отвергается, p-value теста = 0.00000005072026. Следовательно, следует оценивать отдельные модели для разных уровней образования и нельзя оценивать общую модель.

5.8

☼Выберите любые две, желательно незначимых на уровне значимости 10% переменные, и при помощи LR теста или теста Вальда проверьте, можно ли исключить их из модели: при этом необходимо формально записать нулевую и альтернативную гипотезы, статистику теста и её распределение.

Ответ. Неограниченная модель:

$$\mathbb{P}\{marriage_i\} = F(\beta_0 + \beta_1 age_i + \beta_2 age_i^2 + \beta_3 work_i + \beta_4 workxage_i + \varepsilon_i)$$

Ограниченная модель:

$$\mathbb{P}\{marriage_i\} = F(\beta_0 + \beta_1 age_i + \beta_2 age_i^2 + \varepsilon_i)$$

$H_0 : \beta_3 = \beta_4 = 0$
$H_A : \beta_3 \neq 0$ or/and $\beta_4 \neq 0$
Статистика теста:

$$LR = 2(lnL_{UR} - lnL_R) \sim \chi_2^2$$

Две степени свободи хи-квадрата, так как два линейно независимых ограничения в нулевой гипотезе.

p-value теста очень близко к 0. Расчёты выдают 0. Следовательно, нельзя исключать указанный набор переменных, и гипотеза о том, что они в совокупности равны нулю отвергается. Необходимо использовать неограниченную модель.

6

6.1

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

Пусть x_i - вектор характеристик i-го индивида ($Intercept\ Age\ Age^2\ Work\ Workxage$), а β - вектор коэффициентов. y_i - бинарная переменная, семейное положение индивида. Функция правдоподобия:

$$L = \prod_i \left[F\left(x_i'\beta\right)\right]^{y_i} \left[1 - F\left(x_t'\beta\right)\right]^{1-y_i}$$

Максимизируемый логарифм функции правдоподобия

$$l = \ln L = \sum_i \left[y_i \ln F\left(x_i'\beta\right) + \left(1 - y_i\right) \ln\left(1 - F\left(x_i'\beta\right)\right)\right]$$

$$F(u) = \frac{e^u}{1 + e^u}$$

Результаты оценки логит-модели представлены в Таблице 17

	Estimate	Std. Error	z value	Pr(>|z|)
(Intercept)	-8.5719	0.5109	-16.78	0.0000
age	0.3016	0.0183	16.50	0.0000
I(age^2)	-0.0023	0.0002	-14.55	0.0000
work	2.9526	0.3333	8.86	0.0000
workxage	-0.0441	0.0068	-6.53	0.0000

Таблица 17: Результаты оценки logit-модели

6.2

☼Проинтерпретируйте значения оценок изменений в отношениях шанса по каждой независимой переменной.

Так как для всех индивидов это изменение различно, посчитаем это изменение на моих характеристиках. В результате вычислений при $\delta = 1$ получим результат, представленный в Таблице 18

	Age	Work
Delta ratio	0.0086	0.2526

Таблица 18: Изменение отношения шансов

Следовательно, при увеличении моего возраста на год, отношение вероятностей изменится в положительную сторону (либо вероятность брака повысится, либо вероятность отсутствия брака понизится, либо и то, и другое одновременно) на 0.8%. Аналогично, если я найду работу, то отношение шансов вырастет на 25%.

7

7.1

☼Оцените систему бинарных уравнений на брак и факт употребления алкоголя, предварительно определив, какая из этих переменных влияет на другую. Результат представьте в форме таблицы (можно, например, использовать выдачу из stata, R или python). Дайте интерпретацию изменениям в оценках коэффициентов по сравнению с обычной пробит моделью

Ответ. Предположим, что факт употребления алкоголя влияет на вероятность вступления в брак. В данном случае можно предположить различную направленность эффектов, так как неизвестна частота употребления алкоголя. Например, можно предположить, что женщин может отпугивать мужчина, который часто употребляет алкоголь. С другой стороны, люди, которые употребляют алкоголь просто от случая к случаю в компании друзей могут просто считать это видом социальной активности, тем самым завязывая новые знакомства и вступая впоследствии в отношения.

Оценим следующую систему уравнений (запишем, для простоты, в латентных переменных)

$$\begin{cases} marriage_i^* = \beta_0 + \beta_1 age_i + \beta_2 age_i^2 + \beta_3 work_i + \beta_4 workxage_i + \gamma alc_i + \varepsilon_{1i} \\ alc_i^* = \alpha_0 + \alpha_1 age_i + \alpha_2 age_i^2 + \alpha_3 work_i + \alpha_4 workxage_i + \varepsilon_{2i} \end{cases}$$

Результаты оценки системы уравнений можно найти в Таблице 19 и Таблице 20.

	Estimate	Std. Error	z value	Pr(>|z|)
(Intercept)	-3.53	0.45	-7.90	0.00
age	0.16	0.01	12.44	0.00
I(age^2)	-0.00	0.00	-12.77	0.00
work	1.65	0.18	9.16	0.00
workxage	-0.02	0.00	-6.72	0.00
alc	-1.35	0.11	-12.59	0.00

Таблица 19: Оценки коэффициентов первого уравнения системы

	Estimate	Std. Error	z value	Pr(>|z|)
(Intercept)	-1.20	0.18	-6.61	0.00
age	0.07	0.01	8.74	0.00
I(age^2)	-0.00	0.00	-8.43	0.00
work	0.95	0.13	7.09	0.00
workxage	-0.01	0.00	-4.47	0.00

Таблица 20: Оценки коэффициентов второго уравнения системы

7.2

☼Проинтерпретируйте значение оценки коэффициента корреляции между случайными ошибками рассматриваемых уравнений.

Ответ. Также был оценён коэффициент корреляции между случайными ошибками (theta):

$$theta = 0.889, CI_theta = (0.523, 0.978)$$

Следует заметить также, что коэффициент γ перед переменной alc в первом уравнении значим. Исходя из этого, а также значимости коэффициента θ можно сделать следующие выводы:

☼ Наличенствует причинно-следственная связь между употреблением алкоголя и вероятностью вступления в брак.

☼ Так как коэффициент корреляции случайных ошибок значим, в первом уравнении будет присутствовать эндогенность. Случайные ошибки, включённые в уравнение латентной переменной *alc*, будут коррелировать с собственными случайными ошибками первого уравнения, что будет порождать эндогенность.

8

8.1

☼Постройте Рок-Кривые для пробит и логит моделей, а также сравните их предсказательную силу. Какую из этих моделей лучше использовать для получения уровня специфичности, равного 0.8? Запишите любой уровень специфичности, при котором лучше использовать пробит модель, чем логит модель, или поясните, почему такого уровня специфичности не существует, по крайней мере если в качестве критерия использовать точность предсказания внутри выборки, на которой осуществлялось оценивание моделей.

Ответ. ROC-кривые представлены на Рис.4.

ROC-AUC hetprobit = 0.7402

ROC-AUC hetlogit = 0.6177

Accuracy hetprobit = Accuracy hetlogit = 0.727

Так как доли правильных ответов равны, сравним их по ROC-AUC. У hetprobit площадь под ROC-кривой больше, следственно, она предпочтительнее. Даже визуально это видно. hetlogit-модель более прижата к оси y = x. Значит, она склоняется в сторону наивной модели.

Рассмотрим уровни sensitivity для specificity в районе 0.8.

Для hetprobit ближайшей точкой будет $(0.462, 0.798)$. Для hetlogit такой точкой будет $(0.392, 0.816)$. Очевидно из графика, что hetlogit при приближении к точке specificity = 0.8 резко не вырастет, так что в данном случае hetprobit-модель позволяет получить больший уровень sensitivity при фиксированной specificity. Следовательно, hetprobit-модель является более предпочтительной.

Одним из таких уровней специфичности, как раз, является 0.8.

8.2

☼Сравните пробит, логит и модель с гетероскедастичной случайной ошибкой по критериям AIC и BIC, выбрав лучшую из них. Поясните, почему вы не можете сравнить эти модели по данному критерию с линейно-вероятностной, по крайней мере используя стандартную выдачу из stata или R

Ответ. Результаты подсчёта информационных критериев представлены в Таблице 21. По обоим информационным критериям побеждает модель с гетероскедастичной случайной ошибкой. Данные модели нельзя сравнить с линейно-вероятностной моделью, так как для неё неизвестно значение правдоподобия. Она оценивается через МНК, в котором не максимизируется правдоподобие. Следовательно, нельзя вычислить для линейной модели информационные критерии.

Модель	AIC	BIC
Probit	4565.54	4597.16
Logit	4566.16	4597.78
Hetprobit	4559.72	4597.67

Таблица 21: Информационные критерии

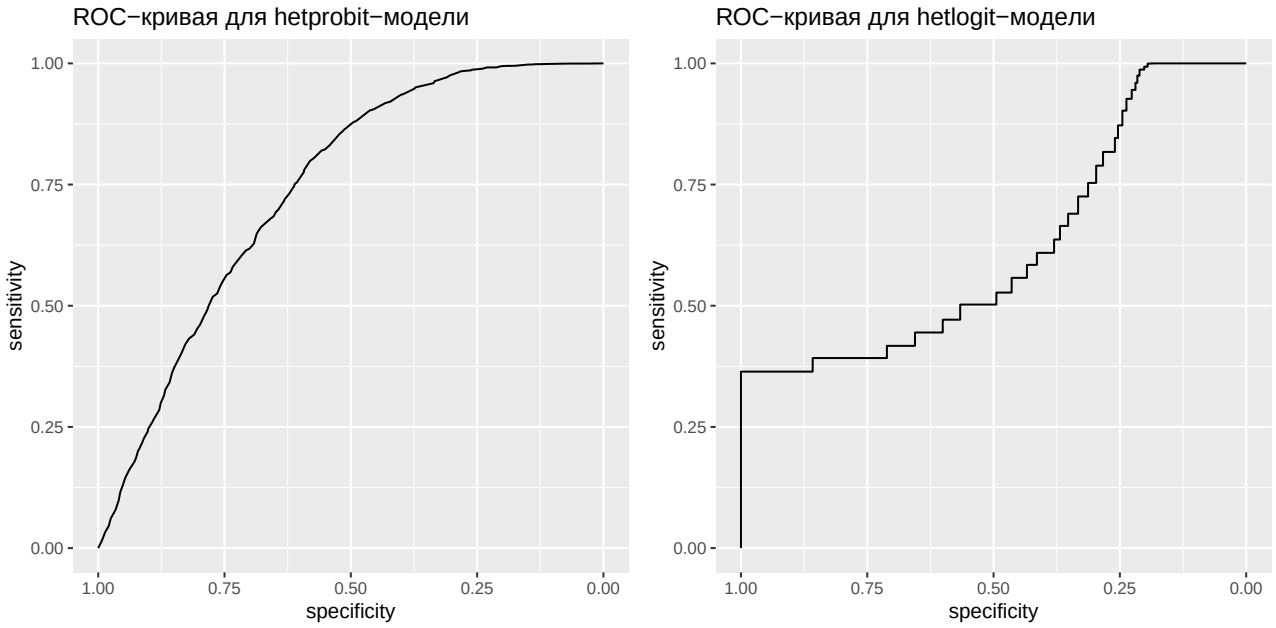

Рис. 4: ROC-кривые

8.3

8.4

Разделим выборку на обучающую и тестовую. В последней оставим 500 наблюдений, в первой - все остальные. Обучим на тренировочнонй выборке три модели и посчитаем MSE на тестовой. Получим Результат, представленный в Таблице 22 . Как можно было ожидать, побеждает модель с гетероскедастичной случайной ошибкой.

	Hetprobit	Probit	Logit	Naive
Accuracy	0.60	0.56	0.58	0.52

Таблица 22: Models accuracy