Регрессионные модели для дихотомических зависимых переменных

А.Р.Бессуднов
(bessudnov@gmail.com)

НИУ ВШЭ, факультет социологии

6 апреля 2012

Первая часть курса

- Описательная статистика и графики
- Линейная регрессия: используется для интервальных зависимых переменных
- Регрессия и корреляция
- Нелинейность и эффекты взаимодействия

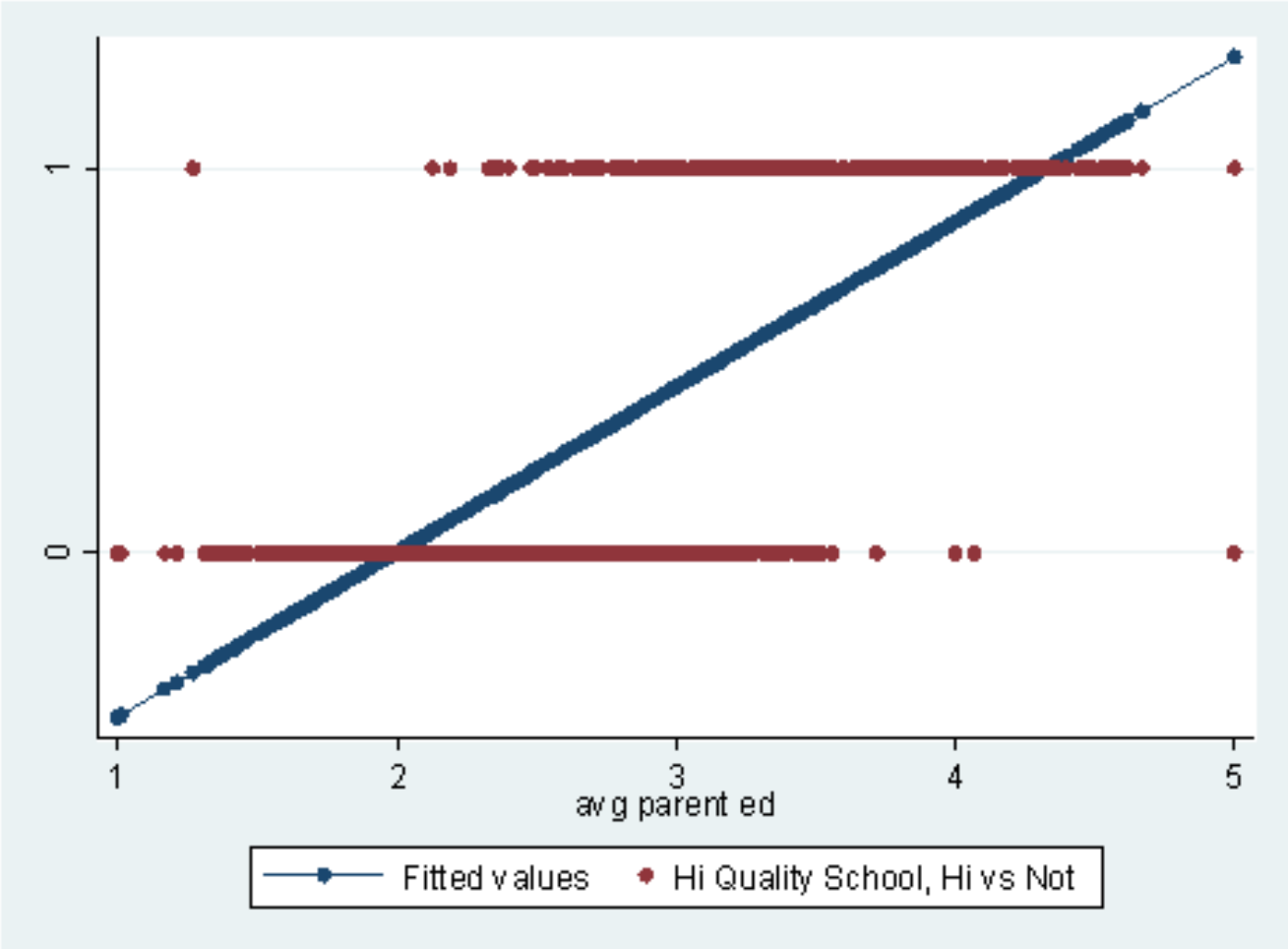
Дихотомические зависимые переменные

- В социальных науках большинство переменных являются категориальными (качественными)
- Во многих случаях в качестве зависимых переменных мы имеем дихотомические переменными; иными словами, мы хотим моделировать связь между предикторами и вероятностью какого-то события или наличия/отсутствия какой-либо характеристики

Линейная модель вероятности

- Наиболее простым решением будет использовать обычную линейную регрессию: $\pi_i = \alpha + \beta X_i$
- Эта модель в применении к дихотомическим зависимым переменным называется линейной моделью вероятности (linear probability model)

Пример: успеваемость и образование родителей в школах Калифорнии



Источник: `http://www.ats.ucla.edu/stat/stata/webbooks/logistic/chapter1/statalog1.htm`

Недостатки линейной модели вероятности

- Остатки распределены не нормально; дисперсия остатков не может быть постоянной
- Предсказанные вероятности могут оказаться за пределами интервала [0,1]

Линейная модель вероятности (2)

- Тем не менее во многих случаях применение линейной модели вероятности оправдано
- Вероятности, меньшие нуля и большие единицы, можно приравнять, соответственно, нулю и единице. Для коррекции стандартных ошибок можно использовать робастную регрессию
- Однако есть другие модели, применение которых в случае дихотомических зависимых переменных считается более оправданным: логит (в основном в социологии, политологии и медицине) и пробит (главным образом в экономике)

Вероятности, шансы и отношения шансов

- Вероятность события – это отношение числа случаев, в которых событие произошло, к общему числу попыток. Монету бросали 100 раз, в 48 случаях выпала решка. Оцененная вероятность выпадения решки 48/100=0.48
- Шансы (odds) – это отношение числа случаев, в которых событие произошло, к числу случаев, в которых событие не произошло. В предыдущем примере шансы равны 48/52=0.92
- Понятно, что вероятность и шансы содержат одну и ту же информацию: $\pi = \frac{odds}{1+odds}$ and $odds = \frac{\pi}{1-\pi}$

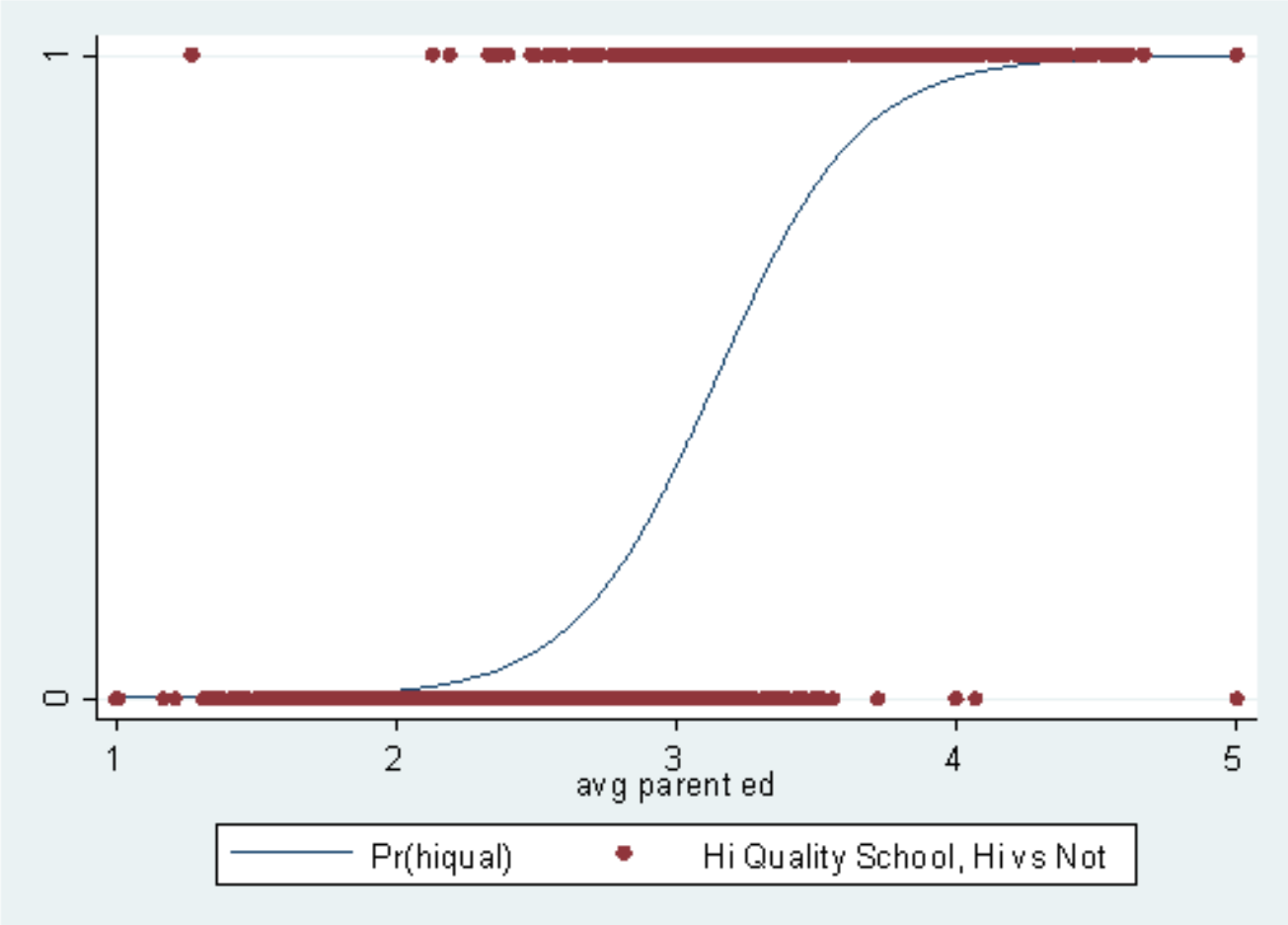
Вероятности, шансы и отношения шансов (2)

- Отношение шансов сравнивает шансы в двух группах. Допустим, что женщины доживают до 70 лет в 8 случаях из 10 – таким образом, их шансы дожить до 70 равны 8/2=4. Допустим, что мужчины доживают до 70 лет в 4 случаях из 10, их шансы дожить до 70 – 4/6=0.33. Отношение шансов (ж/м)=4/0.33=12. Таким образом, шансы женщин дожить до 70 лет в 12 раз выше, чем шансы мужчин
- В вероятностях: вероятность женщин дожить до 70 лет – 0.8, мужчин – 0.4. Таким образом, вероятность женщин дожить до 70 на 40% (или в два раза) выше, чем у мужчин

Логистическая модель

- Логистическая модель преобразует вероятности π таким образом, чтобы они не выходили за пределы интервала [0,1]
- $\text{logit}\pi_i = \log \frac{\pi_i}{1-\pi_i} = \alpha + \beta x$
- $\pi_i = \frac{\exp(\alpha+\beta x)}{1+\exp(\alpha+\beta x)}$

Логистическая кривая



Источник: http://www.ats.ucla.edu/stat/stata/webbooks/logistic/chapter1/statalog1.htm

Интерпретация коэффициентов в логистической модели

- Связь между переменными является положительной, когла $\beta > 0$, и отрицательной, когда $\beta < 0$
- Коэффициенты НЕ могут интерпретироваться непосредственно как изменение вероятности события при изменении независимой переменной, поскольку они представляют изменение $\mathrm{logit}\pi$
- Исходя только из коэффициентов, сложно судить о силе связи
- Возможным решением является преобразование коэффициентов в отношения шансов: $\frac{\pi_i}{1-\pi_i} = \exp(\alpha + \beta x) = e^{\alpha}(e^{\beta x})$
- Тогда на изменение независимой переменной на единицу приходится изменение шансов события в e^{β} раз
- Например, если $e^{\beta} = 1.3$, то шансы события увеличиваются на 30% для каждого изменения x на единицу

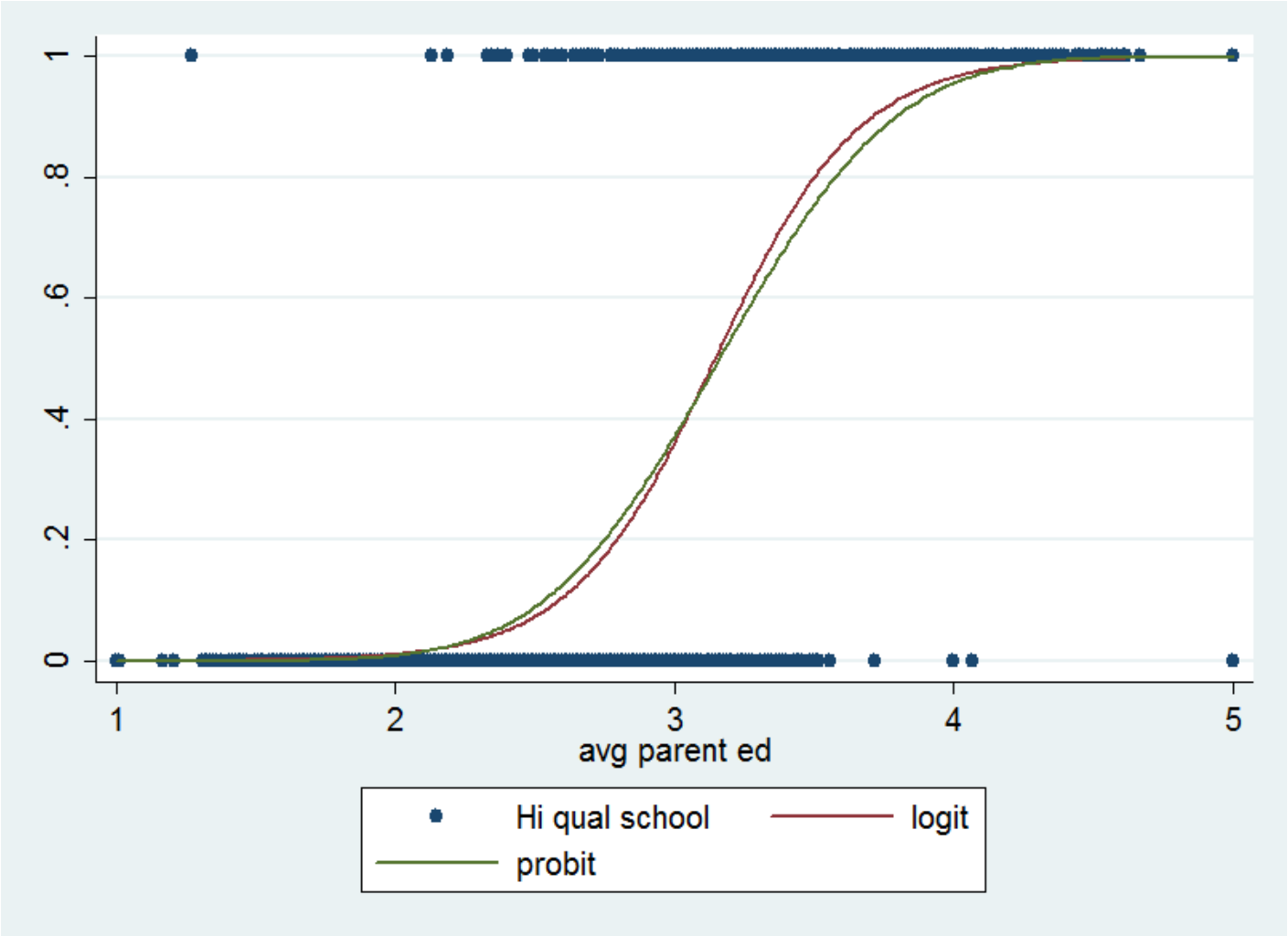
Предсказанные вероятности

- Однако интерпретация логистической регрессии в терминах отношений шансов по-прежнему является не вполне интуитивно понятной
- Наилучшим решением является интерпретация логистической регрессии в терминах предсказанных вероятностей для определенных значений предикторов: $\pi_i = \frac{\exp(\alpha+\beta x)}{1+\exp(\alpha+\beta x)}$

Пробит-модель

- Если для ограничения вероятностей в интервале [0,1] вместо логистической функции накопленного распределения выбрать функцию нормального распределения, то мы получим пробит-модель
- $\pi_i = \Phi(\alpha + \beta X_i) = \frac{1}{\sqrt{2\pi}} \int_{-\infty}^{\alpha+\beta X_i} \exp\left(-\frac{1}{2}Z^2\right) dZ$
- Интерпретация: $\mathrm{probit}\pi$ является значением z, при котором в стандартном нормальном распределении левосторонняя вероятность равна π. probit(0.5)=0, probit(0.95)=1.6, и т.д.
- Коэффициенты в пробит-регрессии нельзя интерпретировать в терминах отношений шансов
- На практике логит и пробит-модели дают очень близкие результаты
- В социологии обычно используется логистическая регрессия

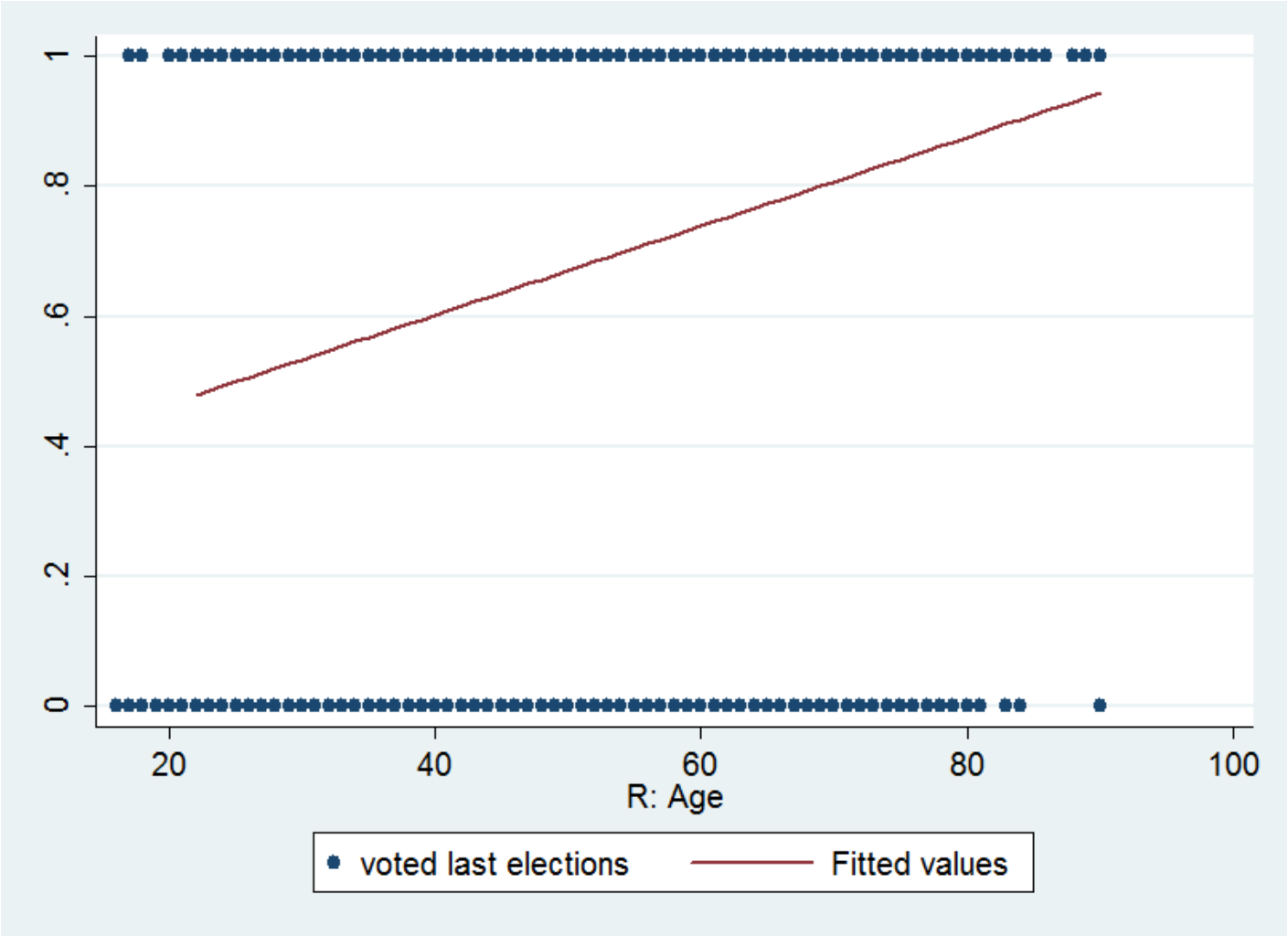
Логит и пробит



Пример 1: вероятность голосования и возраст (ISSP 2003 Россия)

- Линейная модель вероятности: $\text{vote} = 0.11 + 0.01 * \text{age}$

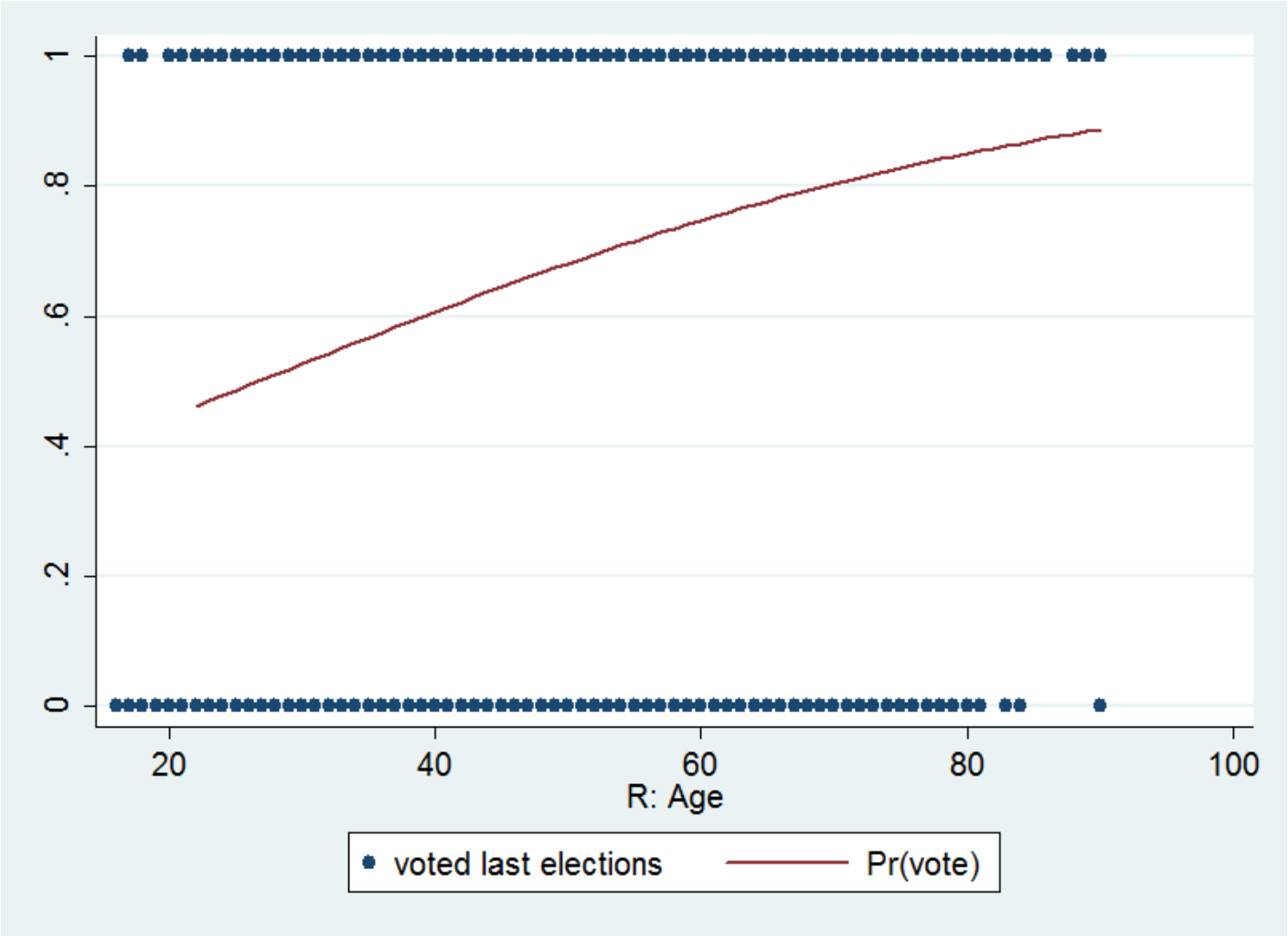
Линейная модель вероятности для голосования



Логистическая модель вероятности голосования

- $\log \frac{\pi_i}{1-\pi_i} = -0.8688 + 0.0326 * \text{age}$
- Вычисляем экспоненту: $\exp(0.0326) = 1.03$
- OR(age)=1.03, т.е. на изменение возраста на один год приходится увеличение шансов голосования на 3%
- Шансы для 40-летнего человека:
  $\frac{\pi_i}{1-\pi_i} = exp(-0.8688 + 0.0326 * 40) = 1.545$
- Рассчитываем вероятность: $\frac{1.545}{1+1.545} = 0.61$
- Для 50-летнего человека:
  $\frac{\pi_i}{1-\pi_i} = exp(-0.8688 + 0.0326 * 50) = 2.14$, вероятность голосования равна $\frac{2.14}{1+2.14} = 0.68$

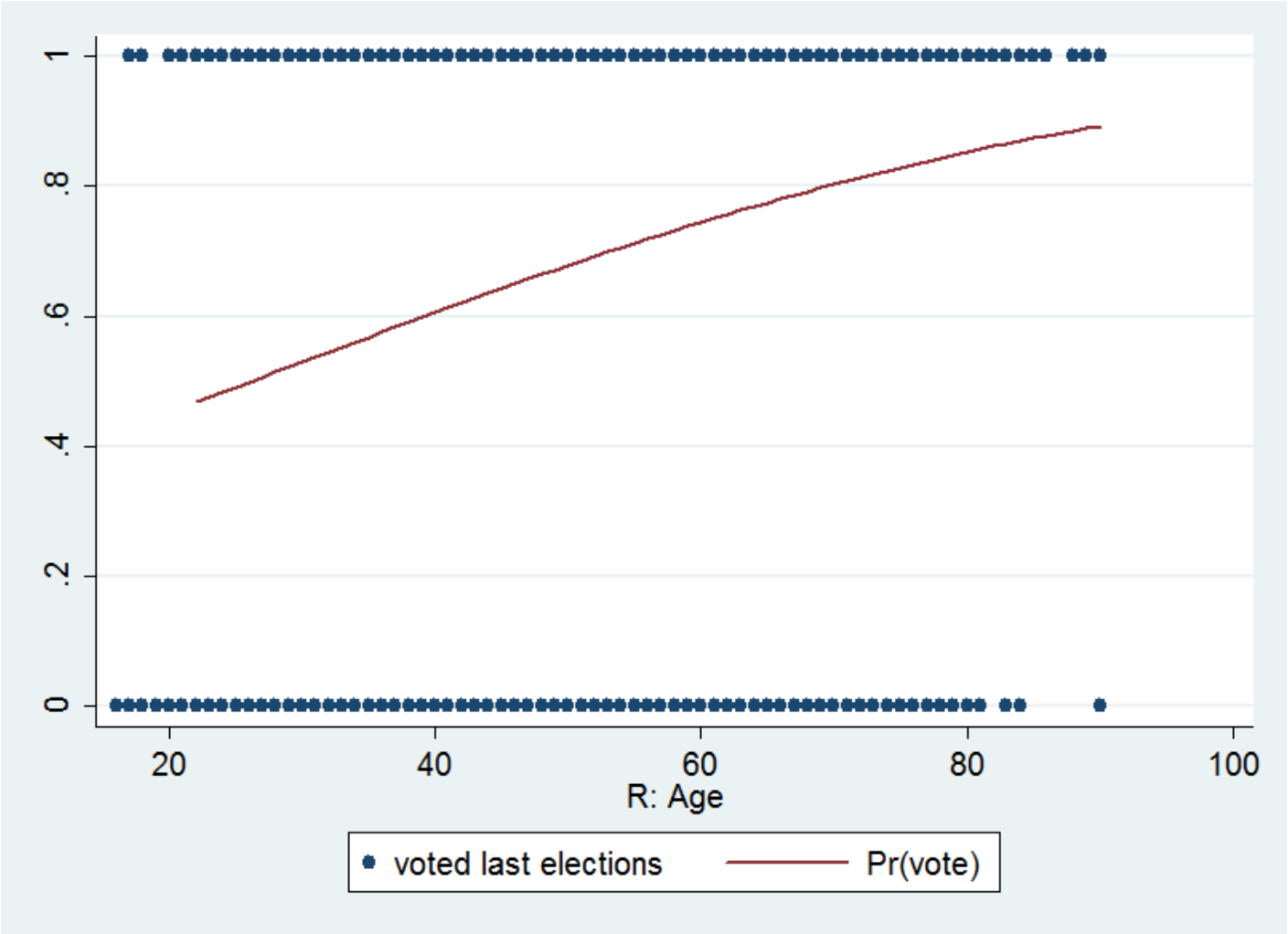
Логистическая модель (2)



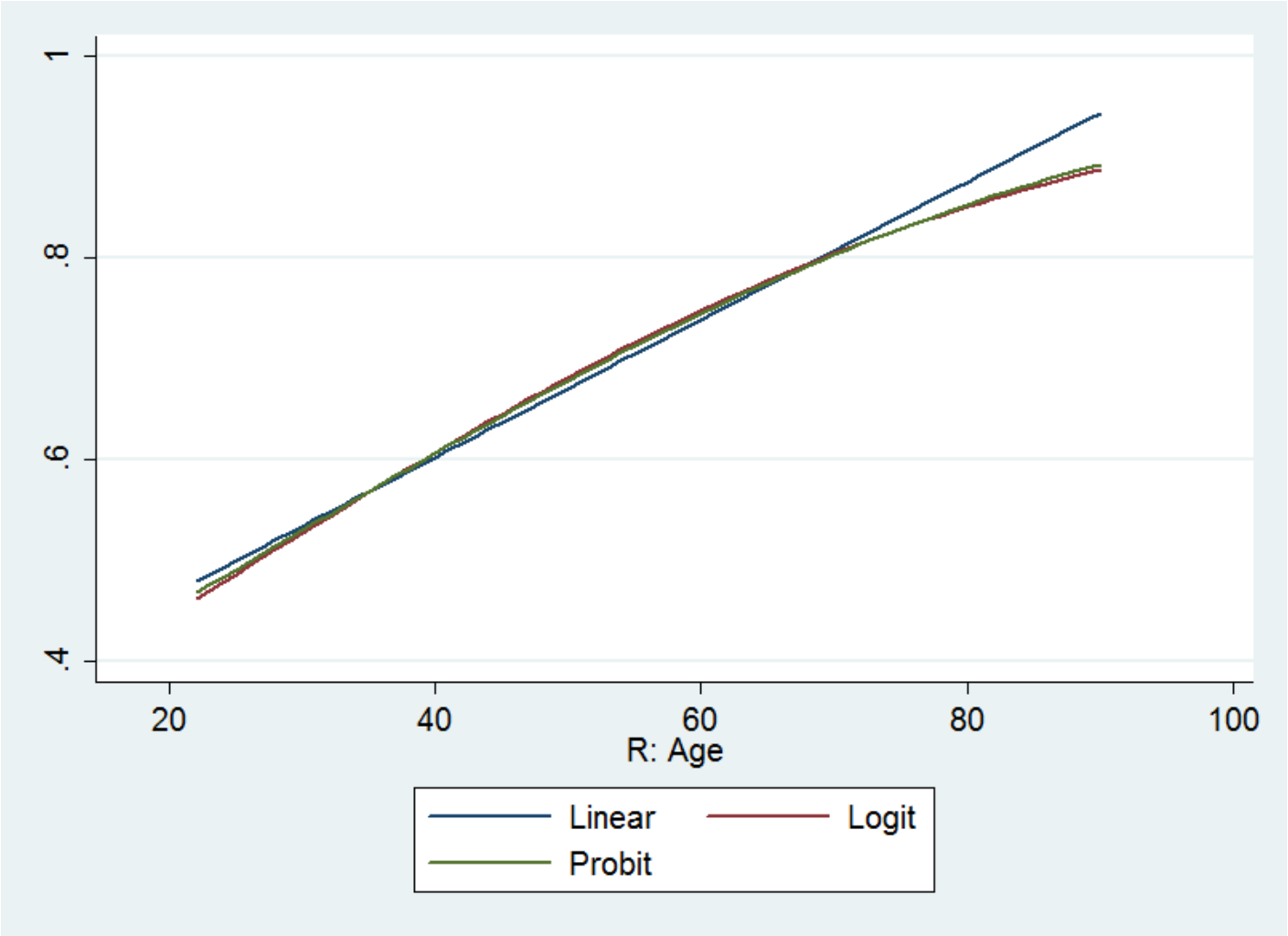
Пробит-модель вероятности голосования

- $\text{probit}\pi = -0.51 + 0.02 * \text{age}$
- Для 40-летнего человека: $\text{probit}\pi = -0.51 + 0.0195 * 40 = 0.27$
- probit(0.27)=0.61

Пробит-модель (2)



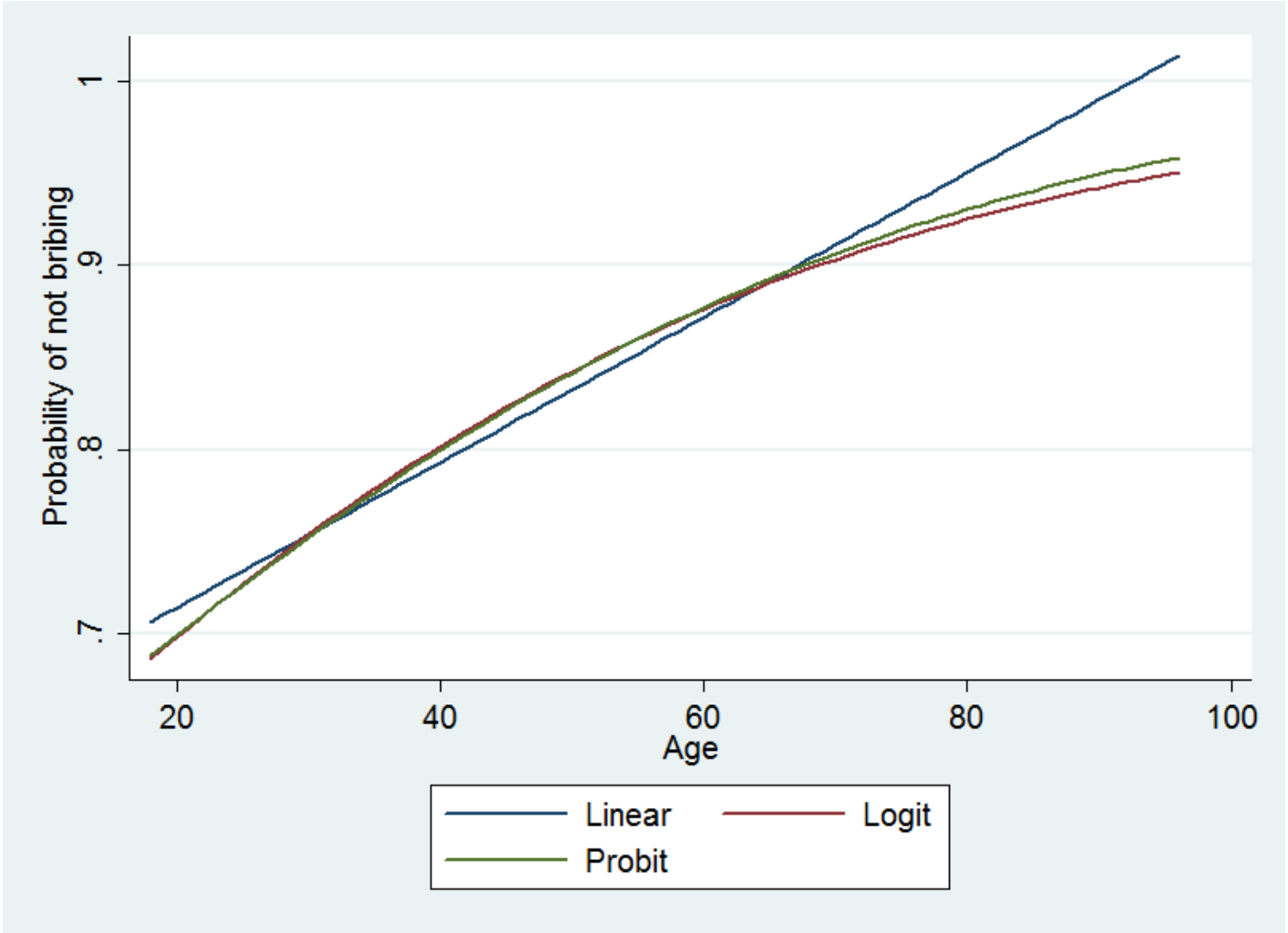
Все модели вместе



Пример 2: вероятность дачи взятки в России и возраст (“Георейтинг”, ФОМ 2003)

- Линейная модель вероятности: $\text{nobribe} = 0.64 + 0.004 * \text{age}$
- Логит: $\text{logit}(\text{nobribe}) = 0.28 + 0.028 * \text{age}$
- Пробит: $\text{probit}(\text{nobribe}) = 0.2 + 0.016 * \text{age}$

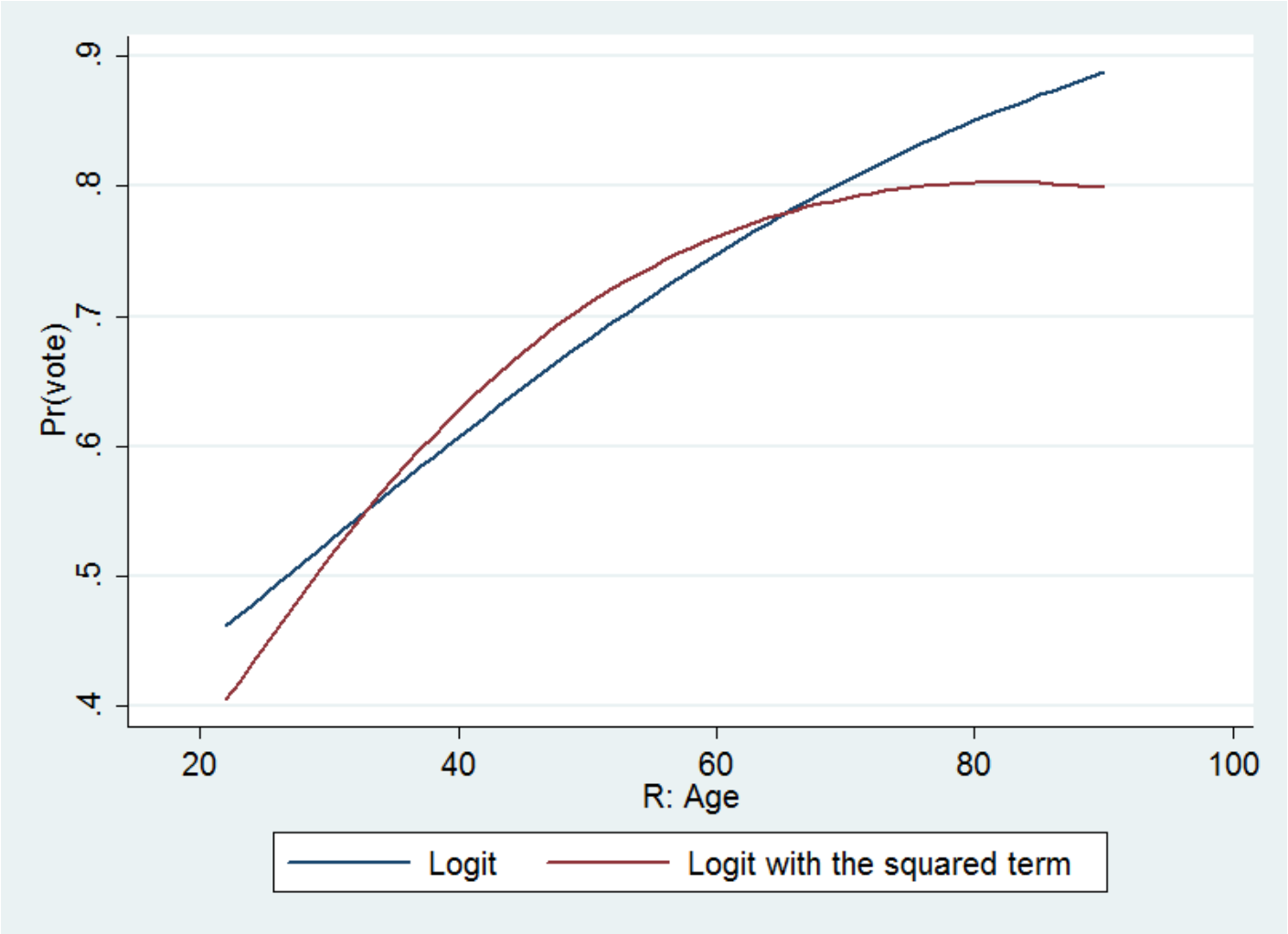
Вероятность дачи взятки и возраст



Преобразования и взаимодействия

- Преобразования переменных могут быть необходимы точно так же, как в линейной регрессии, в случаях, когда мы хотим лучше отразить нелинейность связи между вероятностью события и предиктором
- В логистической регрессии также необходимо учитывать взаимодействие между переменными

Логит-модель и логит-модель с квадратичным членом



Какой метод выбрать?

- Логит и пробит-регрессии дают практически идентичные результаты
- Допустимо также применение линейной модели вероятности, если предсказанные вероятности относительно далеки от 0 и 1

“Качество” модели

- В логит и пробит-регрессии гораздо сложнее измерить пропорцию объясненной дисперсии зависимой переменной
- Существуют несколько разных показателей псевдо-R^2, но к ним следует относиться с осторожностью

Оценка и статистический вывод

- Вместо метода наименьших квадратов (МНК, OLS), для оценки логит- и пробит-моделей используется метод максимизации функции правдоподобия (MLE, maximum likelihood estimation)
- Вкратце, идея заключается в том, чтобы выбрать те параметры модели, при которых вероятность получить имеющиеся данные наиболее велика. См. курс математической статистики
- Чем выше логарифм правдоподобия, тем лучше модель
- Для тестирования гипотез о статистической значимости коэффициентов, используется тест Вальда (Wald test) или тест отношения правдоподобий (likelihood-ratio test)

Заключение

- Для моделирования дихотомических зависимых переменных чаще всего используются логит или пробит-модели. Их результаты удобнее всего интерпретировать в терминах предсказанных вероятностей
- Во многих случаях допустимо применение линейной модели вероятности