Порядковая и мультиномиальная логистическая регрессия

А.Р.Бессуднов
(bessudnov@gmail.com)

НИУ ВШЭ, факультет социологии

20 апреля 2012

Порядковая и мультиномиальная логистическая регрессия

- Порядковая логистическая регрессия применяется тогда, когда зависимая переменная является порядковой, т.е. когда категории упорядочены, но расстояние между ними неизвестно (пример - шкала Лайкерта).
- Мультиномиальная логистическая регрессия применяется тогда, когда зависимая переменная является номинальной, т.е. мы имеем несколько неупорядоченных категорий (пример - поддержка партий на выборах).
- Оба метода являются “надстройкой” над обычной логистической регрессией, однако имеют свои допущения и требуют навыка интерпретации результатов.
- Как и в случае с регрессией с дихотомической зависимой переменной, вместо логита может использоваться пробит, со схожими результатами.

Логистическая регрессия в терминах скрытой переменной

- Вернемся к случаю простой парной логистической регрессии. Ее можно представить в следующей форме:
  $y_i^* = \alpha + \beta x_i + \epsilon_i$, где y^* - скрытая зависимая переменная.
- Мы, однако, наблюдаем только $y = 0$ или $y = 1$:
  $$y_i = \begin{cases} 1 & y_i^* > 0 \\ 0 & y_i^* \leq 0 \end{cases}$$
- Мы хотим оценить вероятность того, что $y = 1$ на определенном уровне x:
  $Pr(y = 1|x) = Pr(y^* > 0|x)$

Логистическая регрессия в терминах скрытой переменной (2)

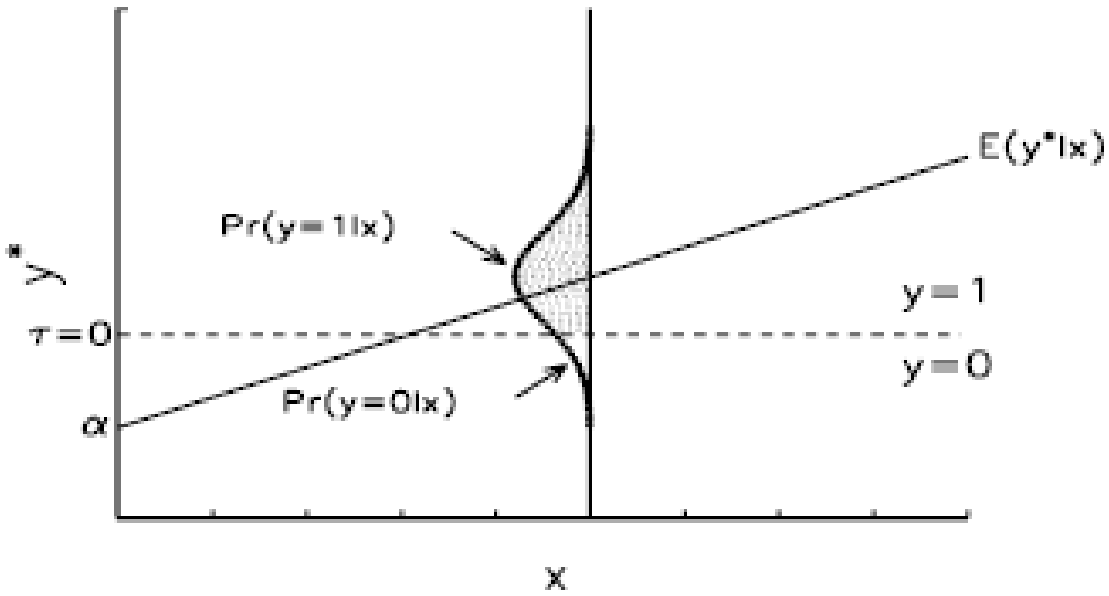


Figure 4.1: Relationship between latent variable y^* and $\Pr(y = 1)$ for the BRM.

(Источник: J.Scott Long and J.Freese, "Regression Models for Categorical Dependent Variables Using Stata")

Логистическая регрессия в терминах скрытой переменной (3)

- $Pr(y = 1|x) = Pr(y^* > 0|x)$ можно переписать как:
  $Pr(y = 1|x) = Pr(\epsilon > -[\alpha + \beta x]|x)$
- Таким образом, $Pr(y = 1|x)$ зависит от распределения ошибки ϵ. В логистической модели ошибка распределена логистически с дисперсией $\pi^2/3$, в пробит-модели - нормально с дисперсией 1.
- Отсюда, в логистической регрессии:

$$Pr(y = 1|x) = \frac{\exp(\alpha + \beta x)}{1 + \exp(\alpha + \beta x)}$$

Порядковая логистическая регрессия в терминах скрытой переменной

- Теперь представим, что вместо двух значений 0 и 1 ненаблюдаемой переменной y^* мы наблюдаем несколько значений, скажем $y = 1, 2, 3, 4$, но не знаем расстояние между этими категориями

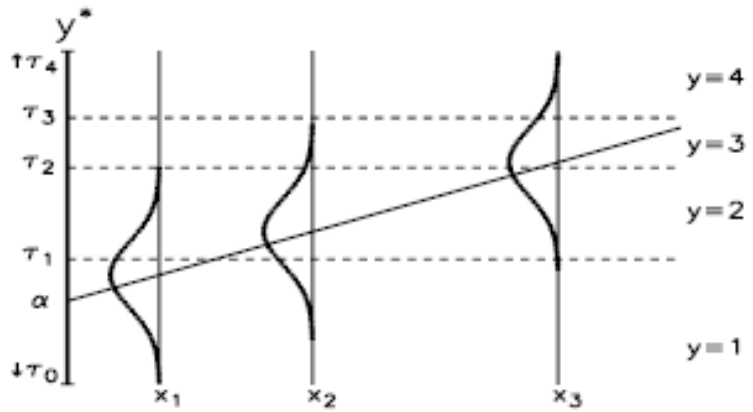


Figure 5.1: Relationship between observed y and latent y^* in ordinal regression model with a single independent variable.

(Источник: J.Scott Long and J.Freese, "Regression Models for Categorical Dependent Variables Using Stata")

Порядковая логистическая регрессия в терминах скрытой переменной (2)

- $y_i = \begin{cases} 1 & \tau_0 = -\infty \leq y_i^* < \tau_1 \\ 2 & \tau_1 \leq y_i^* < \tau_2 \\ 3 & \tau_2 \leq y_i^* < \tau_3 \\ 4 & \tau_3 \leq y_i^* < \tau_4 = \infty \end{cases}$
- Мы хотим оценить вероятность $y = 1, 2, 3, 4$ на определенном уровне x. Сумма этих вероятностей будет равна 1.

Порядковая логистическая регрессия в терминах скрытой переменной (3)

- Вероятность того, что $y = 3$ на определенном уровне x будет равна вероятности того, что значение ненаблюдаемой переменной y^* на этом уровне x больше или равно значения, в котором $y = 2$ и меньше значения, в котором $y = 4$. В математической записи:
  $Pr(y = 3|x) = Pr(\tau_2 \leq y^* < \tau_4|x)$
  $Pr(y = 3|x) = F(\tau_3 - \beta x) - F(\tau_2 - \beta x)$,
  где F - функция распределения (cdf) остатков ϵ.
- $\tau_{1,2,3}$ - это точки (cut-points), которые нам нужно идентифицировать. Заметьте, что для порядковой переменной с 4-мя значениями таких точек 3.

Иными словами:

- В обычной логистической регрессии для каждого наблюдения (случая, респондента) мы можем рассчитать вероятность того, что зависимая переменная y равна 0 и 1, при определенных значениях предикторов.
- В порядковой логистической регрессии все то же самое: мы рассчитываем вероятности того, что зависимая переменная y принимает значения 1,2,k при определенных значениях предикторов.
- Cut-points ограничивают значения линейной функции y^* от предикторов, при которых наблюдаемая переменная y будет принимать те или иные значения. Иными словами, cut-points задают дистанцию между категориями.

Допущение о параллельных регрессионных линиях (parallel regression assumption)

- Порядковую регрессию с числом категорий k можно представить в виде k-1 бинарных логистических регрессий.
- Какова вероятность $y \leq 1$ по сравнению с $y > 1$, $y \leq 2$ по сравнению с $y > 2$, и т.д.
- В порядковой логистической регрессии мы допускаем, что все эти регрессии различаются лишь константой, а коэффициент β остается одинаковым.
- Другое название этого допущения - proportional odds assumption. Если изменение независимой переменной на 1 увеличивает шансы быть в категории 1 в n раз, то оно также увеличивает шансы быть в категории 2 или ниже в n раз, и т.д.

Интерпретация коэффициентов

- Значимый положительный коэффициент: увеличение независимой переменной увеличивает шансы быть в более высокой категории.
- Значимый отрицательный коэффициент: наоборот.
- Более удобно интерпретировать коэффициенты как вероятности, но, как и в случае с обычной логистической регрессией, это требует некоторых преобразований

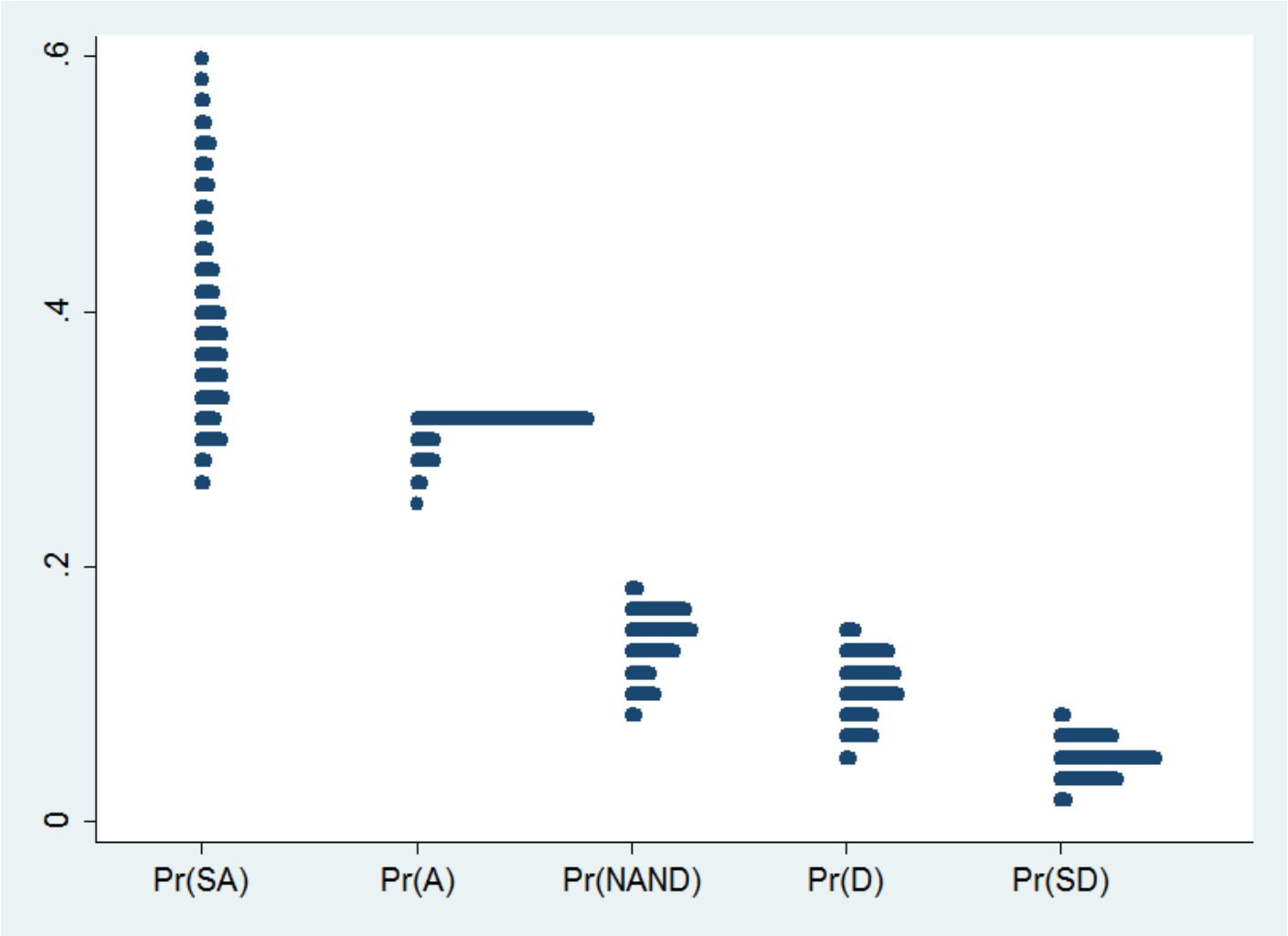
Пример

- Зависимая переменная: “Иммигранты увеличивают уровень преступности” (1 “Полностью согласен”, 2 “Скорее согласен”, 3 “Ни согласен, ни не согласен”, 4 “Скорее не согласен”, 5 “Совершенно не согласен”), ISSP 2003 Россия
- Независимые переменные: пол, возраст, образование

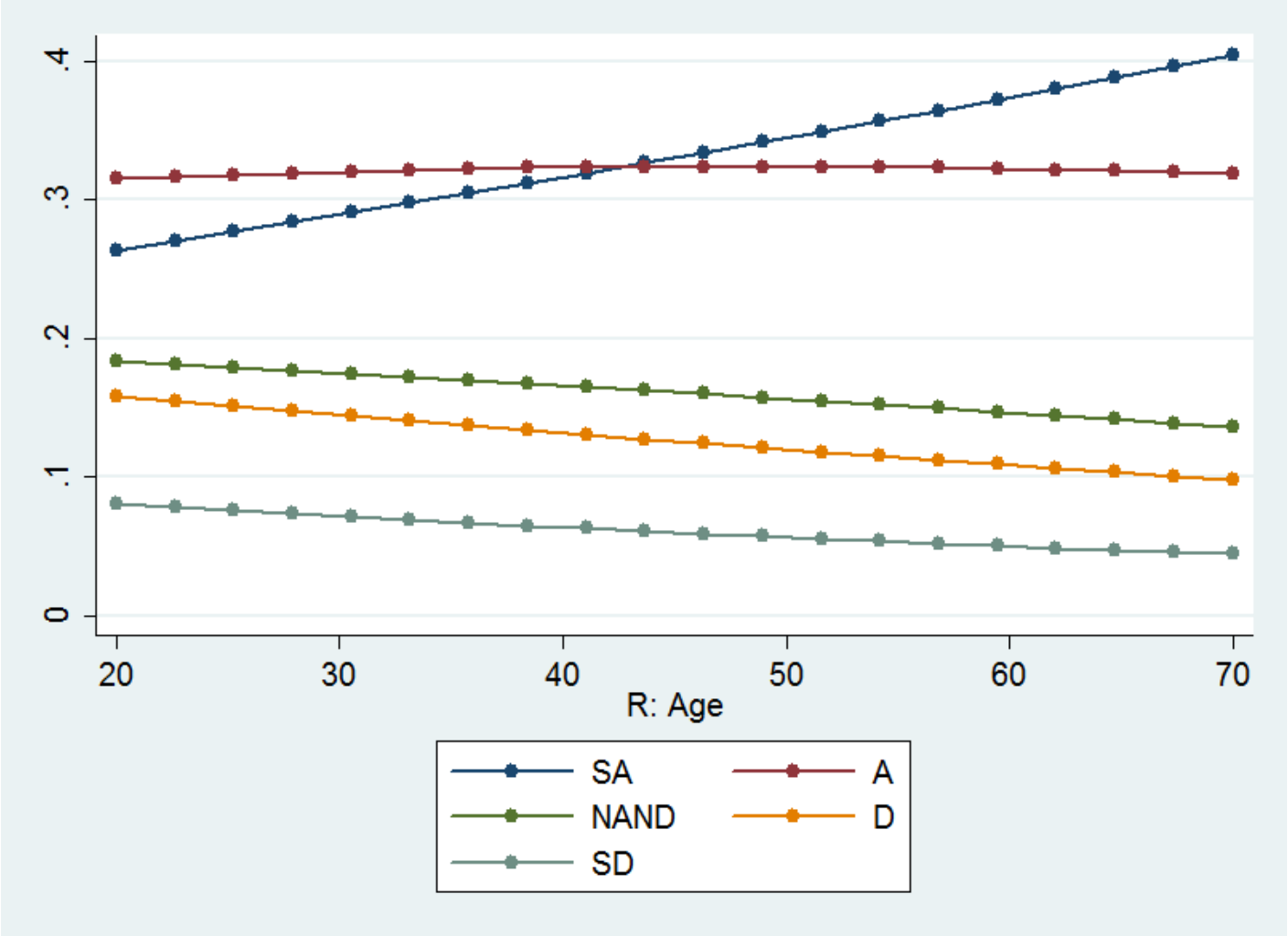
Результаты

переменные	коэф.	ст.ош.	p
мужчины	-0.14	0.08	0.1
возраст	-0.01	0.002	< 0.01
Образование (баз. ниже среднего)			
Законченное среднее	0.25	0.13	0.06
Среднее специальное	0.29	0.13	0.02
Высшее	0.46	0.13	< 0.01
Cut-points			
τ_1	-0.82	0.18	
τ_2	0.52	0.18	
τ_3	1.37	0.18	
τ_4	2.64	0.20	
n	2,080		

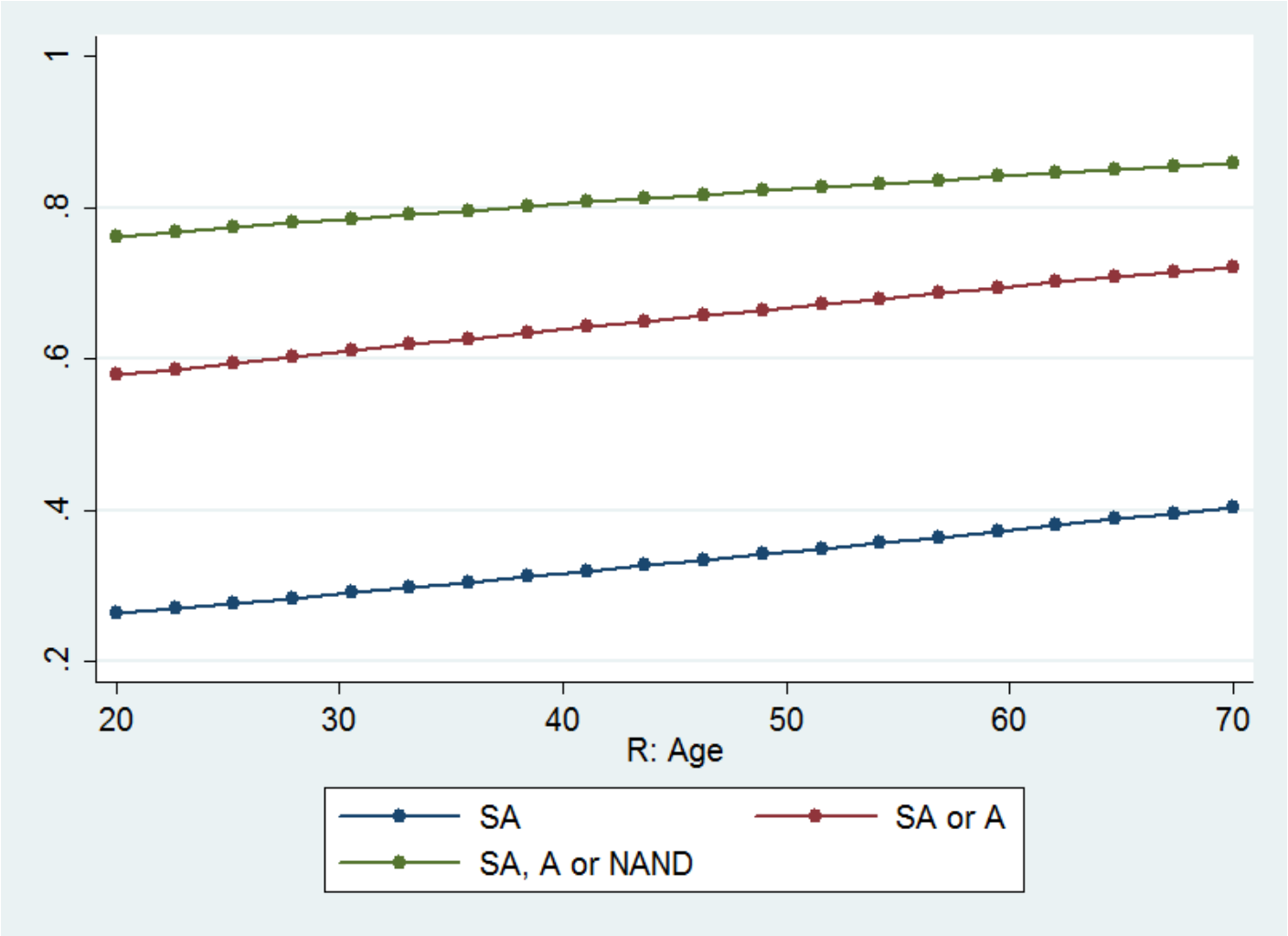
Предсказанные вероятности



Предсказанные вероятности для женщин с высшим образованием



Накопленные вероятности



Результаты линейной регрессии

переменные	коэф.	ст.ош.	p
мужчины	-0.07	0.05	0.2
возраст	-0.01	0.002	< 0.01
Образование (баз. ниже среднего)			
Законченное среднее	0.14	0.08	0.09
Среднее специальное	0.15	0.08	0.06
Высшее	0.24	0.08	< 0.01
Константа	2.34	0.11	< 0.01
n	2,080		

Мультиномиальная логистическая регрессия

- Значения некоторых категориальных зависимых переменных нельзя выстроить в иерархическом порядке (например, партийная поддержка)
- Допустим, у нас есть три партии: консерваторы, либералы и социалисты
- С помощью трех бинарных логистических регрессий мы можем оценить вероятности поддержки консерваторов vs. либералов, консерваторов vs. социалистов, либералов vs. социалистов. Одна из этих моделей является излишней
- По сути, мультиномиальная логистическая регрессия оценивает эти три регрессии одновременно (вводя некоторые ограничения)
- Мультиномиальная логистическая регрессия может использоваться с порядковыми зависимыми переменными, когда не выполняется условие о параллельных регрессионных линиях
- Мультиномиальный логит и пробит дают схожие результаты, в социологии обычно используется мультиномиальный логит

Партийная поддержка как функция образования

- $$\log\left\{\frac{Pr(C|x)}{Pr(L|x)}\right\} = \alpha_1 + \beta_1 ed$$
$$\log\left\{\frac{Pr(C|x)}{Pr(S|x)}\right\} = \alpha_2 + \beta_2 ed$$
$$\log\left\{\frac{Pr(L|x)}{Pr(S|x)}\right\} = \alpha_3 + \beta_3 ed$$
- Одно из уравнений является излишним и может быть выведено из других двух
- Одна из категорий зависимой переменной выбирается как базовая категория, все другие категории сравниваются с ней

В терминах вероятностей

- $$Pr(y = S|ed) = \frac{\exp(\alpha_1 + \beta_1 ed_{S|C})}{1 + \exp(\alpha_1 + \beta_1 ed_{S|C}) + \exp(\alpha_2 + \beta_2 ed_{L|C})}$$

$$Pr(y = L|ed) = \frac{\exp(\alpha_2 + \beta_2 ed_{L|C})}{1 + \exp(\alpha_1 + \beta_1 ed_{S|C}) + \exp(\alpha_2 + \beta_2 ed_{L|C})}$$

$$Pr(y = C|ed) = \frac{1}{1 + \exp(\alpha_1 + \beta_1 ed_{S|C}) + \exp(\alpha_2 + \beta_2 ed_{L|C})}$$

- Консерваторы являются базовой категорией. Сумма вероятностей поддержки трех партий равна единице

Интерпретация модели

- Мы оцениваем вероятность попадания в категорию y по сравнению с базовой категории, в зависимости от предикторов в модели
- Не так важно, какая именно категория выбирается как базовая. Выбор базовой категории зависит от удобства интерпретации
- Сложность интерпретации модели заключается в том, что обычно число параметров велико и их трудно интерпретировать. Например, для зависимой переменной с пятью категориями и регрессии с двумя предикторами, мы получаем 12 параметров (больше, если мы хотим сравнить все категории зависимой переменной попарно)

Интерпретация модели (2)

- β интерпретируется как относительный риск (relative risk) (отношение вероятностей, шансы) оказаться в категории y по сравнению с базовой категорией
- Если коэффициент β положительный и статистически значимый, это означает, что более высокие значения предиктора статистическим значимо связаны с большей вероятностью попасть в категорию y по сравнению с базовой категорией
- Если коэффициент β отрицательный и статистически значимый, это означает, что более высокие значения предиктора статистическим значимо связаны с меньшей вероятностью попасть в категорию y по сравнению с базовой категорией
- Для большей наглядности следует интерпретировать модель графически в терминах предсказанных вероятностей

Допущение о IIA

- IIA означает независимость нерелевантных альтернатив (independence of irrelevant alternatives)
- Допущением модели является это, что отношение вероятностей получить категорию S по сравнению с C не зависит от наличия в модели категории L
- Иными словами, вероятность поддержать социалистов, а не консерваторов, не зависит от того, есть ли в выборке сторонники либералов
- Это допущение можно тестировать
- В мультиномиальной пробит-модели это допущение не делается

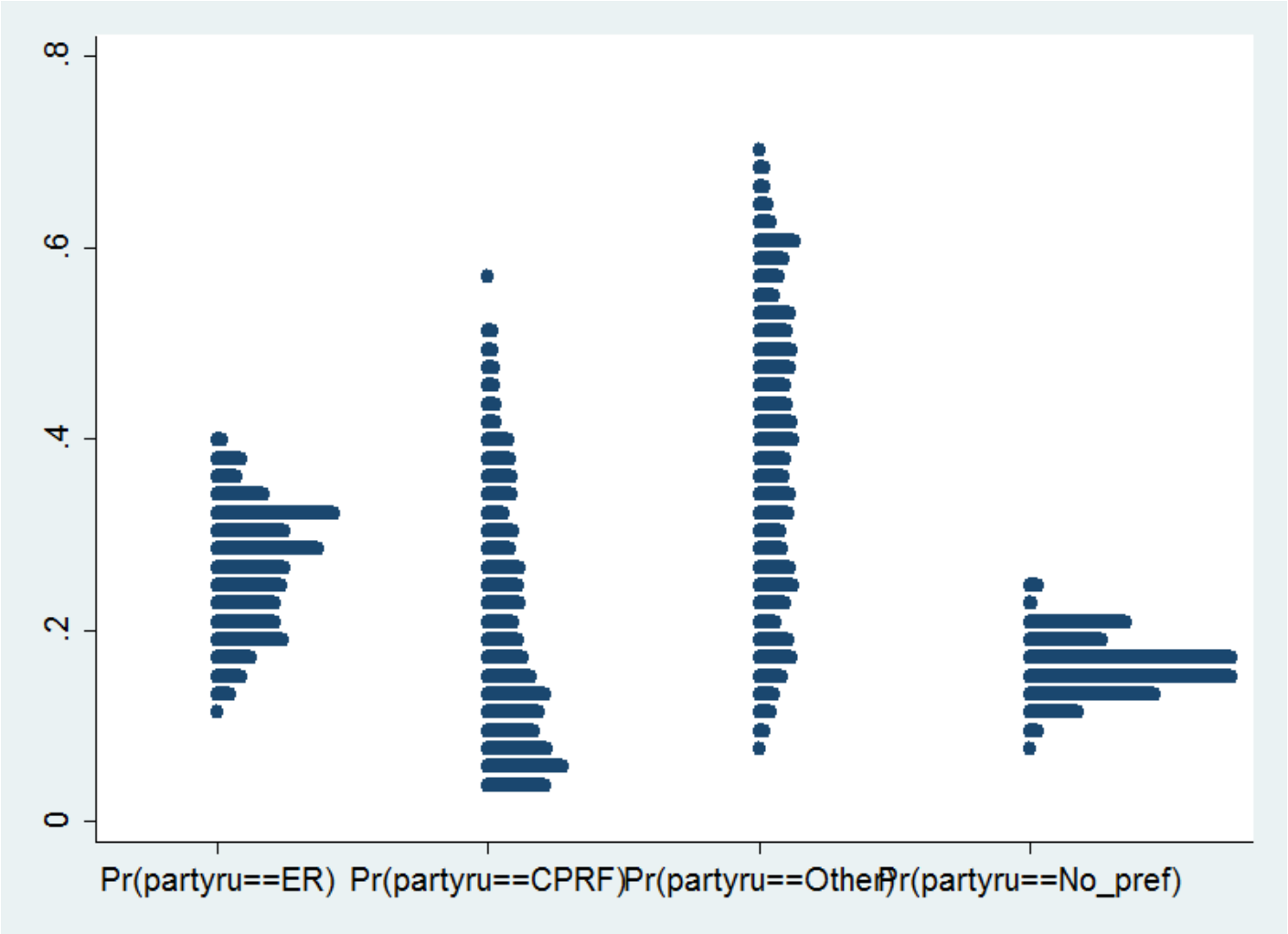
Преобразования и взаимодействия

- Преобразования переменных и эффекты взаимодействия применяется в порядковой и мультиномиальной регрессии так же, как в линейных и бинарных логистических моделях

Пример: поддержка партий как функция возраста, пола и образования

- Зависимая переменная: поддержка партий (1 “Единая Россия”, 2 “КПРФ”, 3 “Другая”, 4 “Нет предпочтений”), ISSP 2003 Россия
- Независимые переменные: пол, возраст, образование
- Базовой категорией является “Единая Россия”
- В результате оценки мультиномиальной логистической регрессии мы получаем набор параметров (константа, коэффициенты для пола, возраста и трех фиктивных переменных для образования) для каждой из трех категорий vs. “Единая Россия”. Всего мы получаем 18 коэффициентов

Предсказанные вероятности



Предсказанные вероятности для мужчин с высшим образованием: зависимость партийной поддержки от возраста