Линейная регрессия: статистический вывод, трансформации переменных и взаимодействие между переменными

А.Р.Бессуднов
(bessudnov@gmail.com)

НИУ ВШЭ, факультет социологии

2 марта 2012

Прошлая лекция

- Модель линейной регрессии используется, когда мы хотим оценить связь одной интервальной зависимой переменной и одной или нескольких независимых переменных. Например: зависимость дохода от пола, образования и места жительства
- $y = a + bx$, где a и b – регрессионные коэффициенты
- Линейная регрессия оценивается методом наименьших квадратов. Регрессионные коэффициенты показывают, на сколько единиц меняется предсказанное значение зависимой переменной, если независимую переменную изменить на единицу (в рамках нашей модели)
- В множественной регрессии коэффициенты при одной переменной интерпретируются с учетом того, что все остальные независимые переменные в модели статистически удерживаются на одном уровне
- Категориальные предикторы (факторы) следует вводить в регрессионное уравнение как (n-1) фиктивных переменных, где n – количество уровней фактора

Сегодня

- Статистический вывод (statistical inference)
- Нелинейные связи и трансформация переменных
- Эффекты взаимодействия

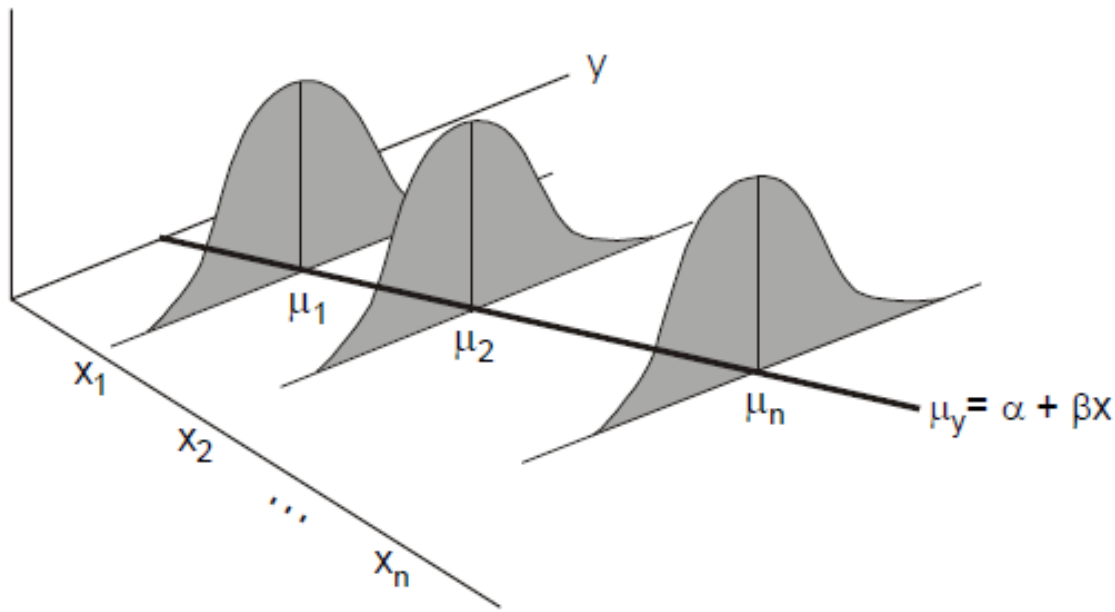
Статистический вывод

- До сих пор мы говорили о регрессии в выборке
- Статистический вывод (statistical inference) – использование информации из случайной выборки для получения представления о генеральной совокупности
- У нас есть статистика, найденная в выборке, мы хотим получить информацию о параметре в генеральной совокупности
- Две формы статистического вывода: доверительные интервалы и тестирование гипотез
- Иными словами, мы хотим узнать, в какой степени регрессионные коэффициенты, полученные на основе анализа выборки, относятся к генеральной совокупности
- Как и в случае средней и пропорции, для того, чтобы делать какие-то выводы о регрессии в генеральной совокупности на основе выборки, выборка должна быть случайной

Условное распределение y

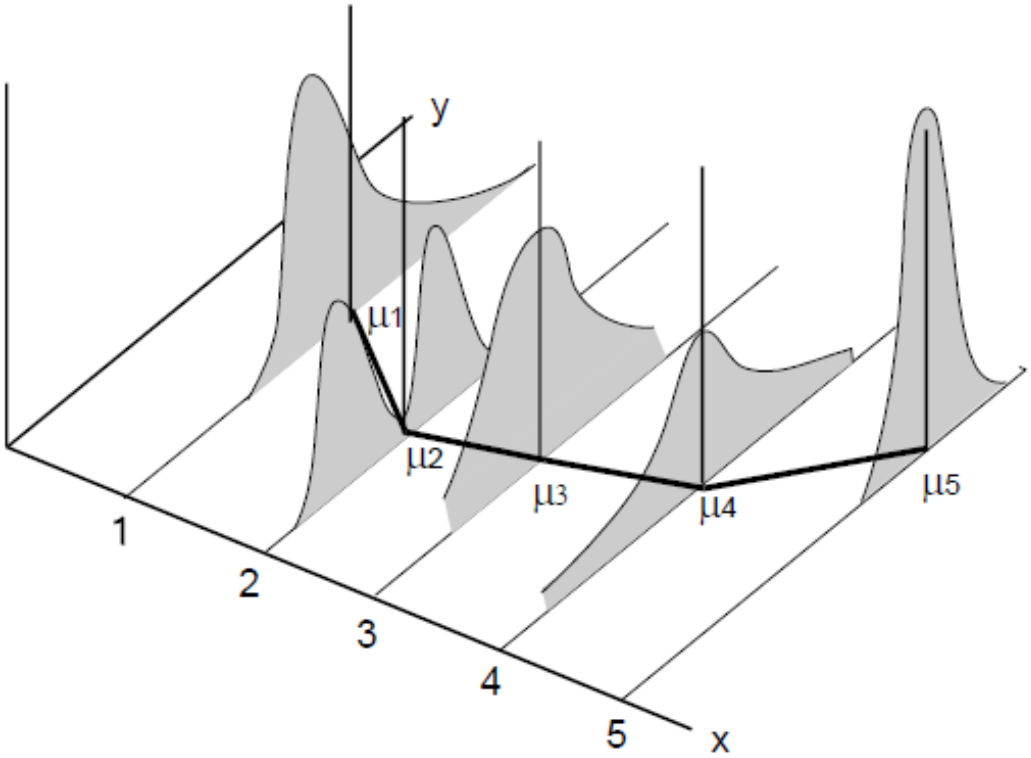
- $y = a + bx$
- Для каждого x существует некоторое распределение y со средней $\mu_y|x$ и стандартным отклонением $\sigma|x$
- Греческими буквами обозначены параметры в генеральной совокупности, латинскими – статистики в выборке
- В МНК-регрессии мы делаем допущение о том, что в генеральной совокупности условное распредение $y|x$ является нормальным и $\sigma|x$ одинаково для всех x. Другими словами, ошибки регрессии распределены нормально с одним и тем же стандартным отклонением
- МНК-регрессия НЕ делает допущения о том, что зависимая переменная должна быть распределена нормально

Иллюстрация



(Courtesy of John Fox)

Это допущения, и они могут быть неверны



(Courtesy of John Fox)

Стандартное отклонение остатков s

- Чтобы оценить, в какой степени регресионная линия в генеральной совокупности может отклоняться от регрессионной линии в выборке, нам нужно знать σ, но это неизвестный нам параметр генеральной совокупности
- Мы можем оценить стандартное отклонение остатков s в выборке
  - $s = \sqrt{\frac{\sum(y-\hat{y})^2}{n-2}}$
  - $n-2$ – количество степеней свободы в модели (n – количество наблюдений, 2 – количество параметров в модели: a и b)

Доверительный интервал для регрессионных коэффициентов

- Нам неизвестен параметр генеральной совокупности β, но мы можем сконструировать доверительный интервал для него, исходя из b
- $\beta \in (b \pm t * SE_b)$, где t – это t-статистика для нужного нам уровня значимости и количества степеней свободы в модели, а SE_b - стандартная ошибка коэффициента b
  - $SE_b = \frac{s}{\sqrt{\sum (x - \bar{x})^2}}$
- Для большого количества степеней свободы t-распределение близко к нормальному
- Таким образом, для $(n > 40)$ и 95%-го доверительного интервала, $\beta \in (b \pm 1.96 * SE_b)$

Интерпретация доверительного интервала

- В случае 95%-го доверительного интервала, исследователь может быть на 95% уверен в том, что искомый параметр попадает в доверительный интервал. Если взять большое количество выборок одного и того же размера, то верный результат будет получен в 95% случаев
- Неверные интерпретации:
  - Существует 95%-я вероятность того, что β находится в доверительном интервале. β либо находится в нем, либо не находится, но нам это неизвестно
  - В 95% случайных выборок b будет находиться в доверительном интервале. Это так для интервала $\beta \pm 1.96 * \sigma$, но нам неизвестны эти параметры генеральной совокупности

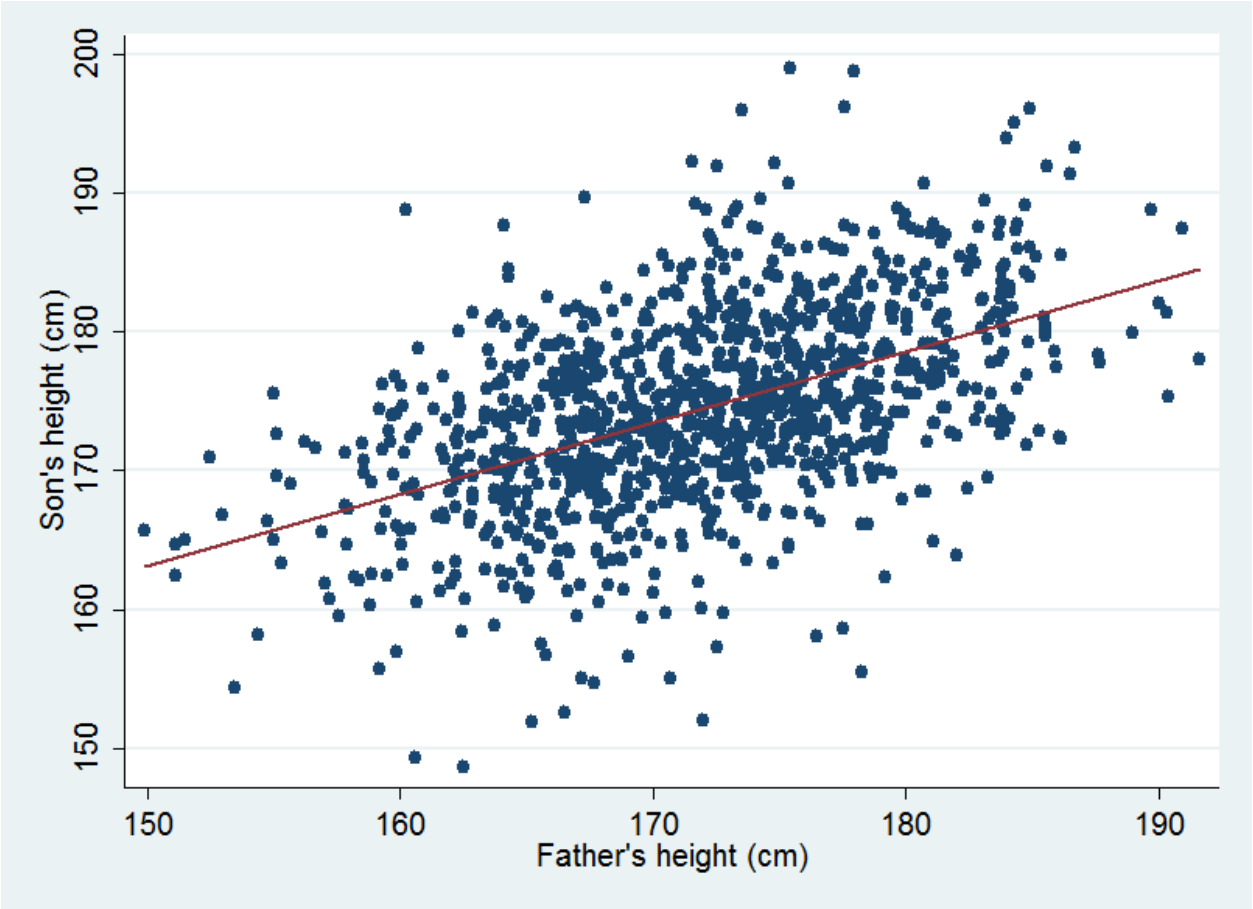
Тест нулевой гипотезы

- Обычно нас интересует нулевая гипотеза о том, что $\beta = 0$
- Альтернативные гипотезы: $\beta \neq 0$ (двусторонняя), $\beta > 0$ or $\beta < 0$ (односторонние)
- Чтобы протестировать нулевая гипотезу, мы можем использовать t-статистику $t = \frac{b}{SE_b}$ с $n - 2$ степенями свободы
- p – вероятность получить t по крайней мере не меньше наблюдаемого, сделав допущение о том, что нулевая гипотеза верна
- Другими словами, мы предполагаем, что в генеральной совокупности связи нет, и смотрим, насколько вероятно получить те данные, что мы имеем, если это так

Тест нулевой гипотезы (2)

- Если p невелика (< 0.05 на 95%-ом уровне значимости), то можно заключить, что вероятность получить такие данные, если нулевая гипотеза верна, мала, и поэтому мы можем отвергнуть нулевую гипотезу
- p не является вероятностью нулевой гипотезы. Подтвердить нулевую гипотезу мы не можем. Мы не можем утверждать, что в генеральной совокупности связь отсутствует, но можем сказать, что у нас недостаточно оснований считать, что она присутствует
- Построение доверительных интервалов и тестирование гипотез связаны между собой. Если ноль попадает в доверительный интервал, то мы не можем отвергнуть нулевую гипотезу

Данные Галтона



Регрессионный вывод

переменные	коэф.	ст.ошибка	t	p
рост отца	0.51	0.03	19	< 0.001
константа	86.1	4.7	18	< 0.001
n	1,078			
R-квадрат	0.25			

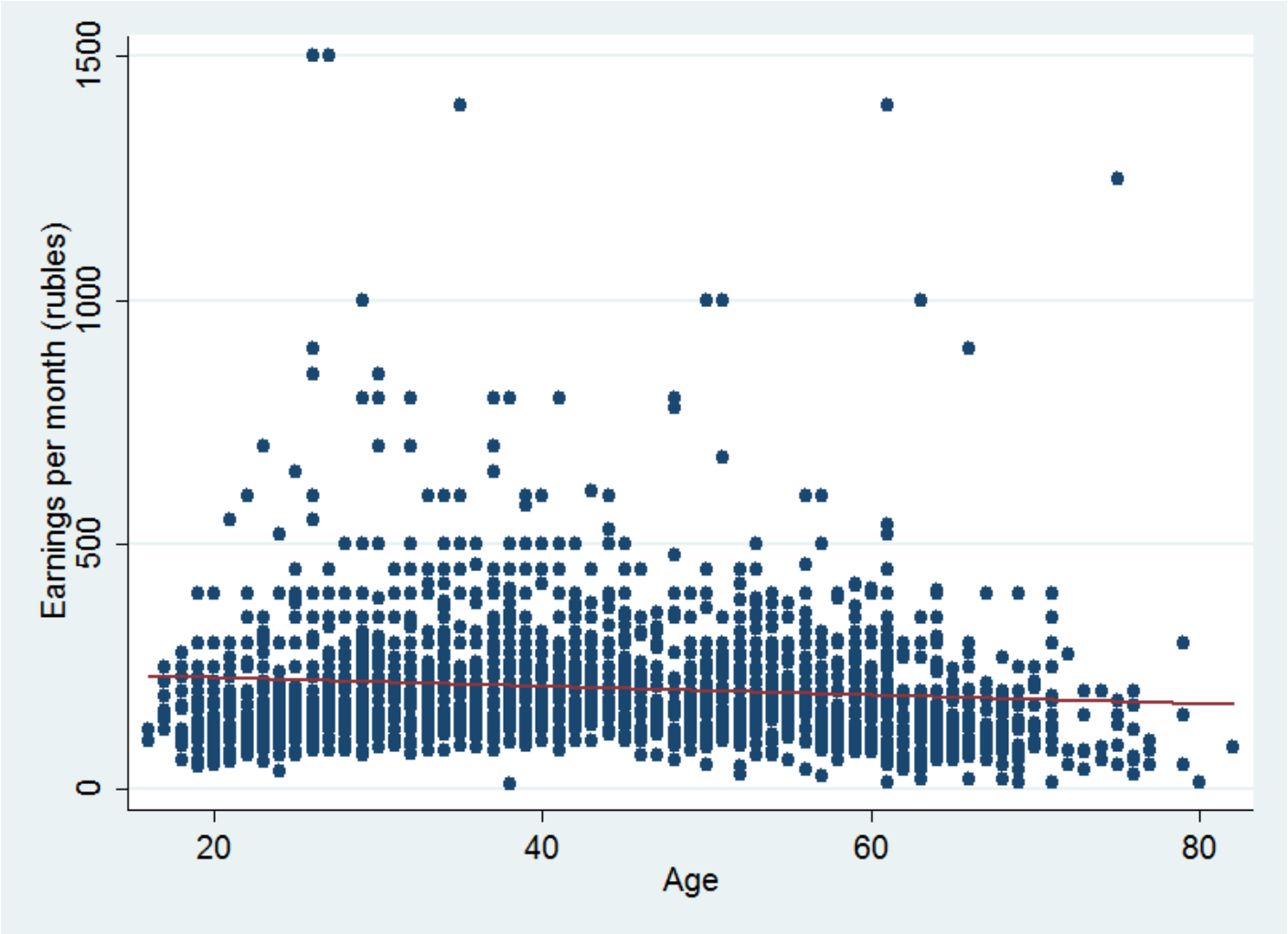
Размер и статистическая значимость регрессионных коэффициентов

- Размер и статистическая значимость коэффициентов – разные вещи, и их не следует путать
- Размер коэффициента b показывает силу эффекта (сколько составляет разница в доходе между людьми с высшим образованием и без него: 100 рублей или 10,000 рублей?)
- Статистическая значимость показывает, насколько мы можем быть уверены в том, что связь, которую мы наблюдаем в выборке, присутствует в генеральной совокупности
- Связь может быть сильной, но статистически не значимой (например, в малых выборках)
- В очень больших выборках практически любой коэффициент будет статистически значим – но это не означает, что эффект важен с практической точки зрения

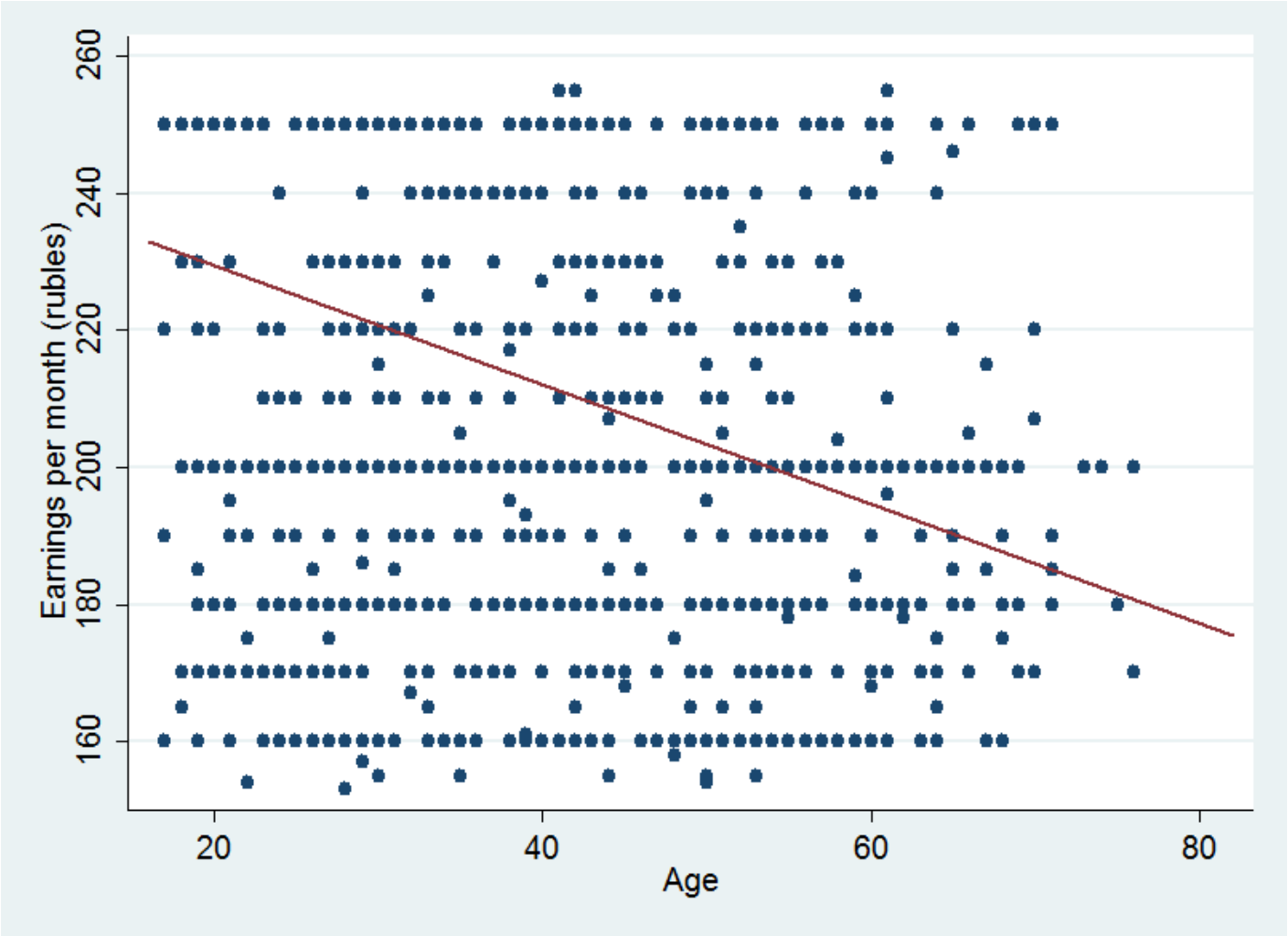
Нелинейные связи

- Одним из допущений линейной регрессии является то, что связь между переменными является линейной
- Часто это не так: например, когда независимой переменной является возраст или доход

Пример: возраст и доход в СССР (1991)



Возраст и доход (2)



Регрессия: доход <- возраст

- Отрицательная связь: чем старше человек, тем меньше доход

переменные	коэф.	ст.ошибка	t	p
возраст (в годах)	-0.87	0.18	-4.8	< 0.001
константа	247	8.4	29	< 0.001
n	2,279			
R-квадрат	0.01			

- Коэффициент для возраста отрицателен и статистически значим. В этой выборке, увеличение возраста на один год связано с уменьшением дохода в среднем на 0.9 рублей в месяц.
  $\text{earnings} = 247 - 0.9 * \text{age}$
- Тем не менее это неверная модель

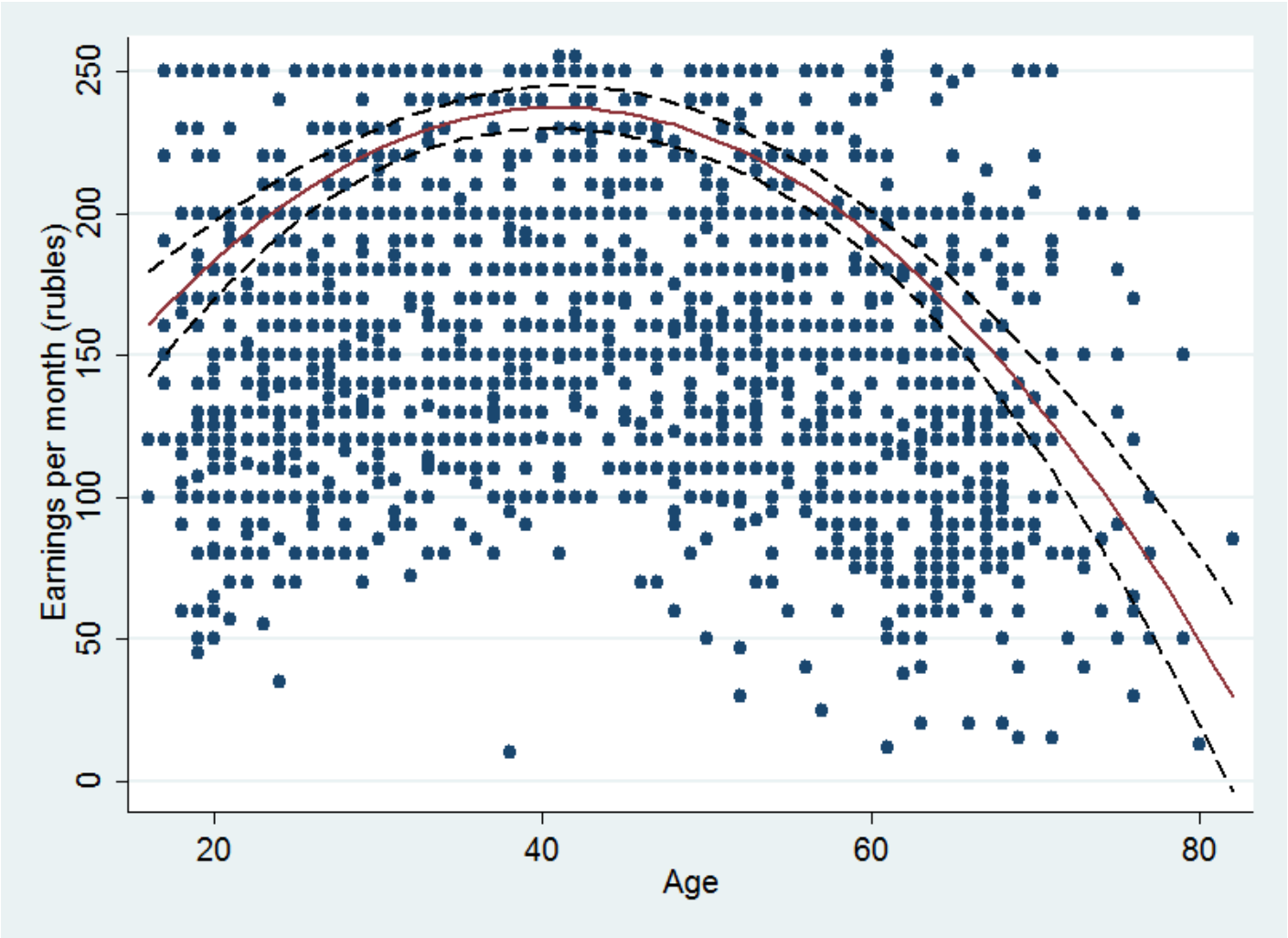
доход <- возраст и возраст в квадрате

- Чтобы оценить нелинейную связь, мы можем добавить в регрессионное уравнение квадратный член для возраста

переменные	коэф.	ст.ошибка	t	p
возраст	10.1	1.1	9	< 0.001
возраст в квадрате	-0.12	0.01	-10	< 0.001
константа	31.5	23.3	1.3	0.18
n	2,279			
R-квадрат	0.05			

- $\text{earnings} = 31.5 + 10.1 * \text{age} - 0.12 * \text{age}^2$
- Отрицательный коэффициент для возраста в квадрате означает, что функция сначала возрастает, а потом убывает. Этот коэффициент статистически значим, что означает, что связь с генеральной совокупности по всей видимости имеет нелинейный характер

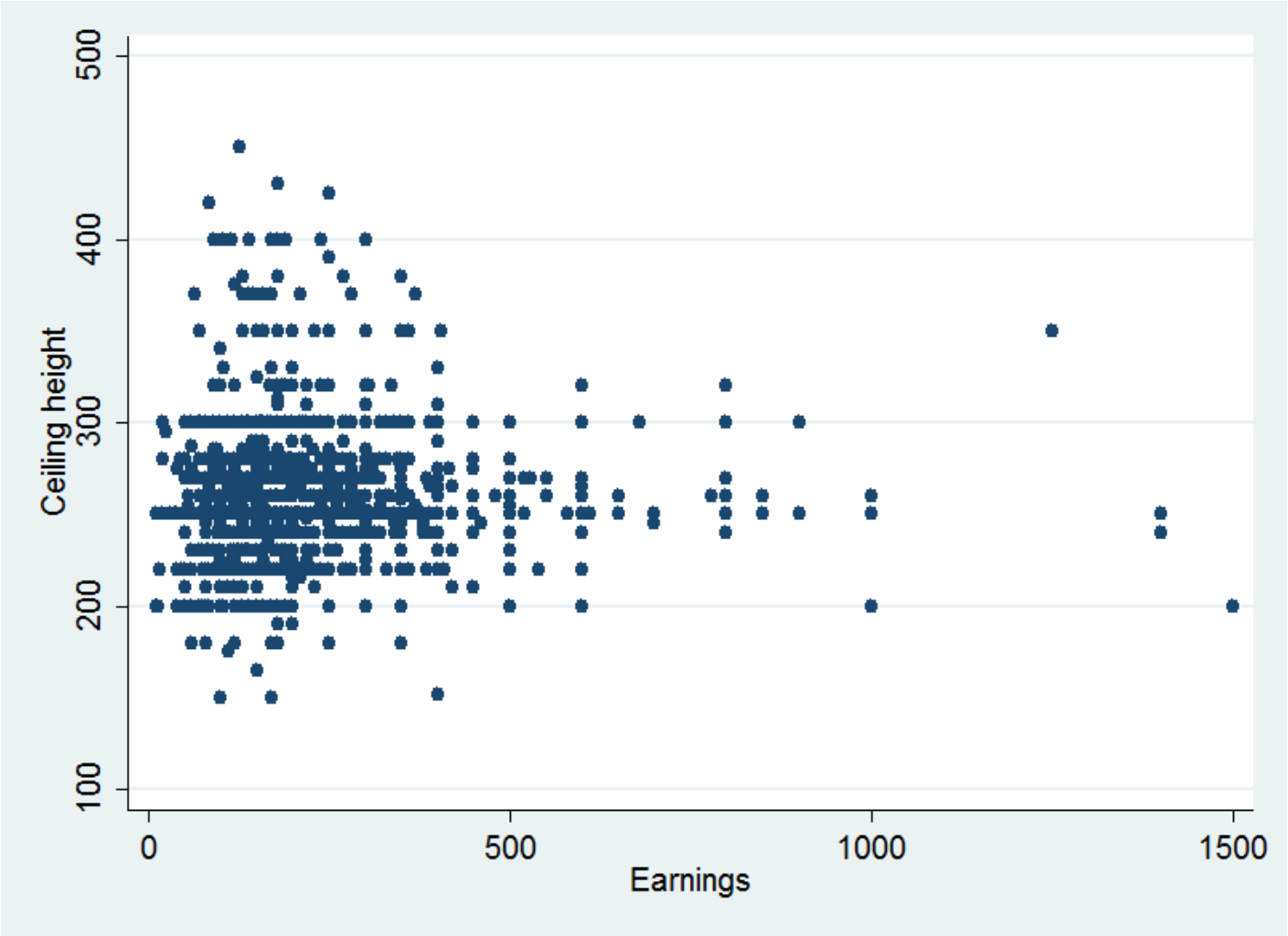
Квадратичная связь между возрастом и доходом (с 95%-м доверительным интервалом)



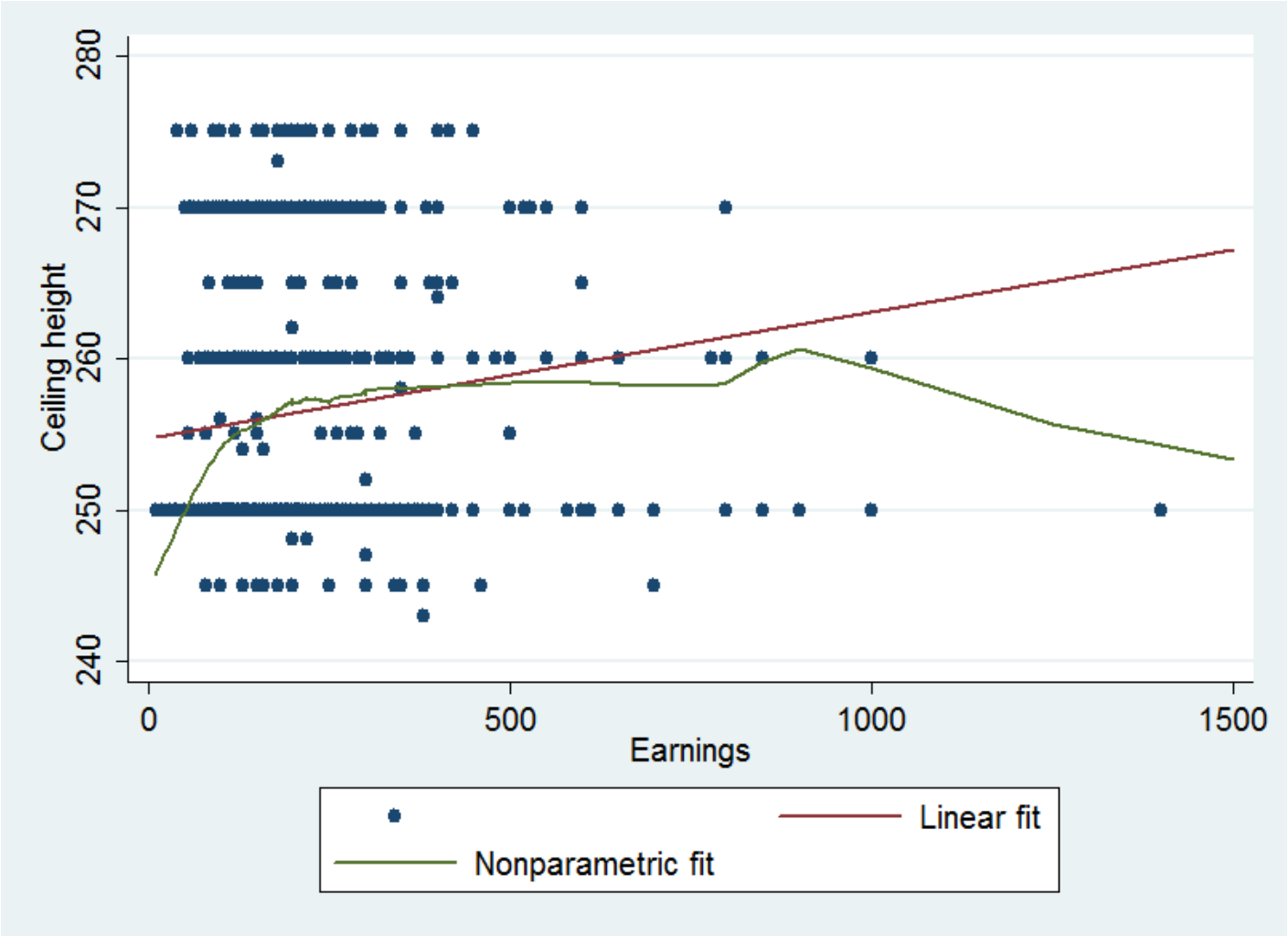
Трансформация независимых переменных

- Иногда требуется трансформировать предикторы таким образом, чтобы превратить нелинейную связь в линейную
- Часто это требуется в случае дохода, который обычно имеет смещенное распределение

Доход и высота потолка (СССР 1991)



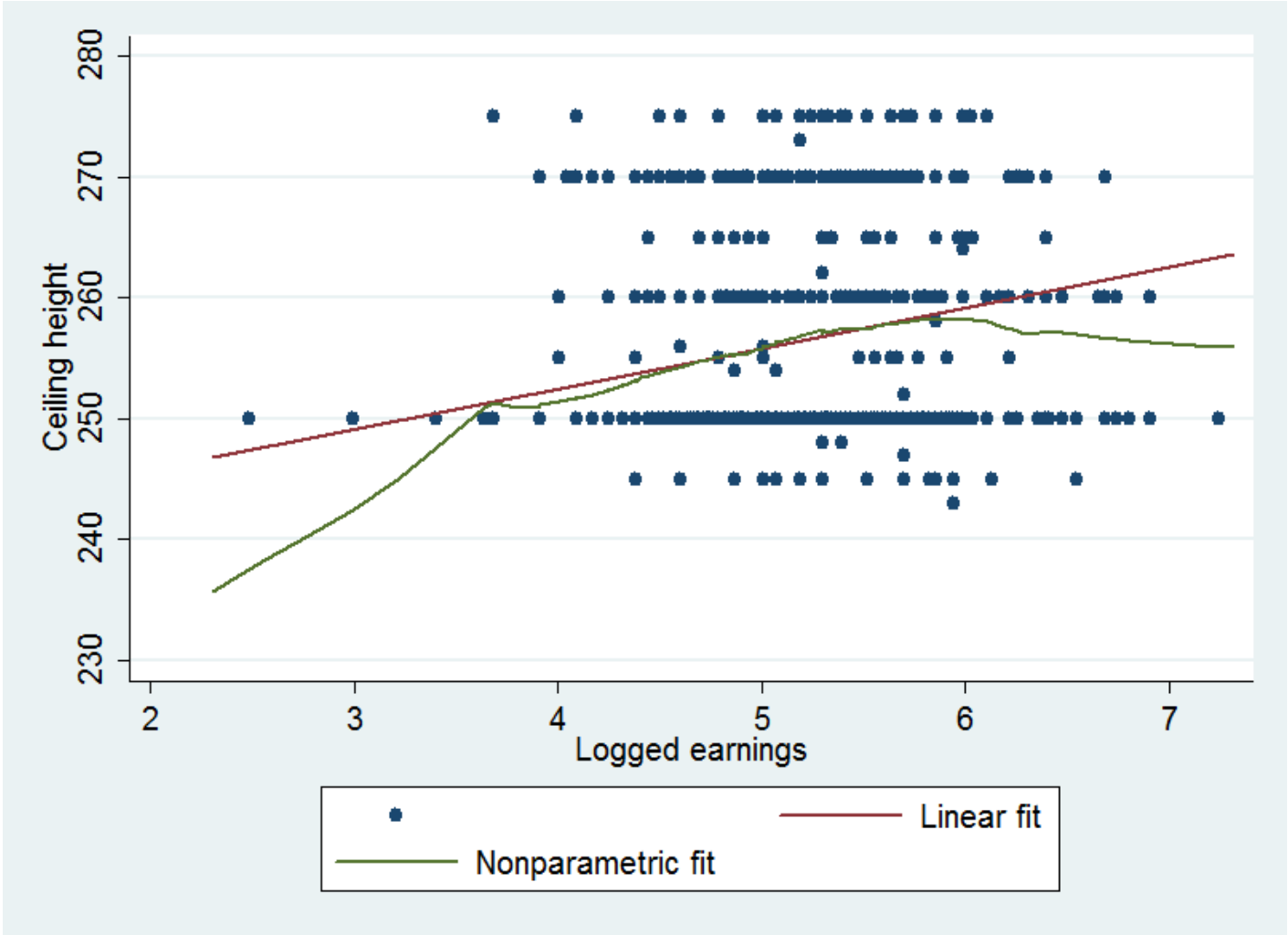
Доход и высота потолка (2)



Высота потолка <- доход

переменные	коэф.	ст.ошибка	t	p
доход	0.008	0.005	1.7	0.09
константа	255	1.2	213	< 0.001
n	2,171			
R-квадрат	0.001			

Логарифм дохода



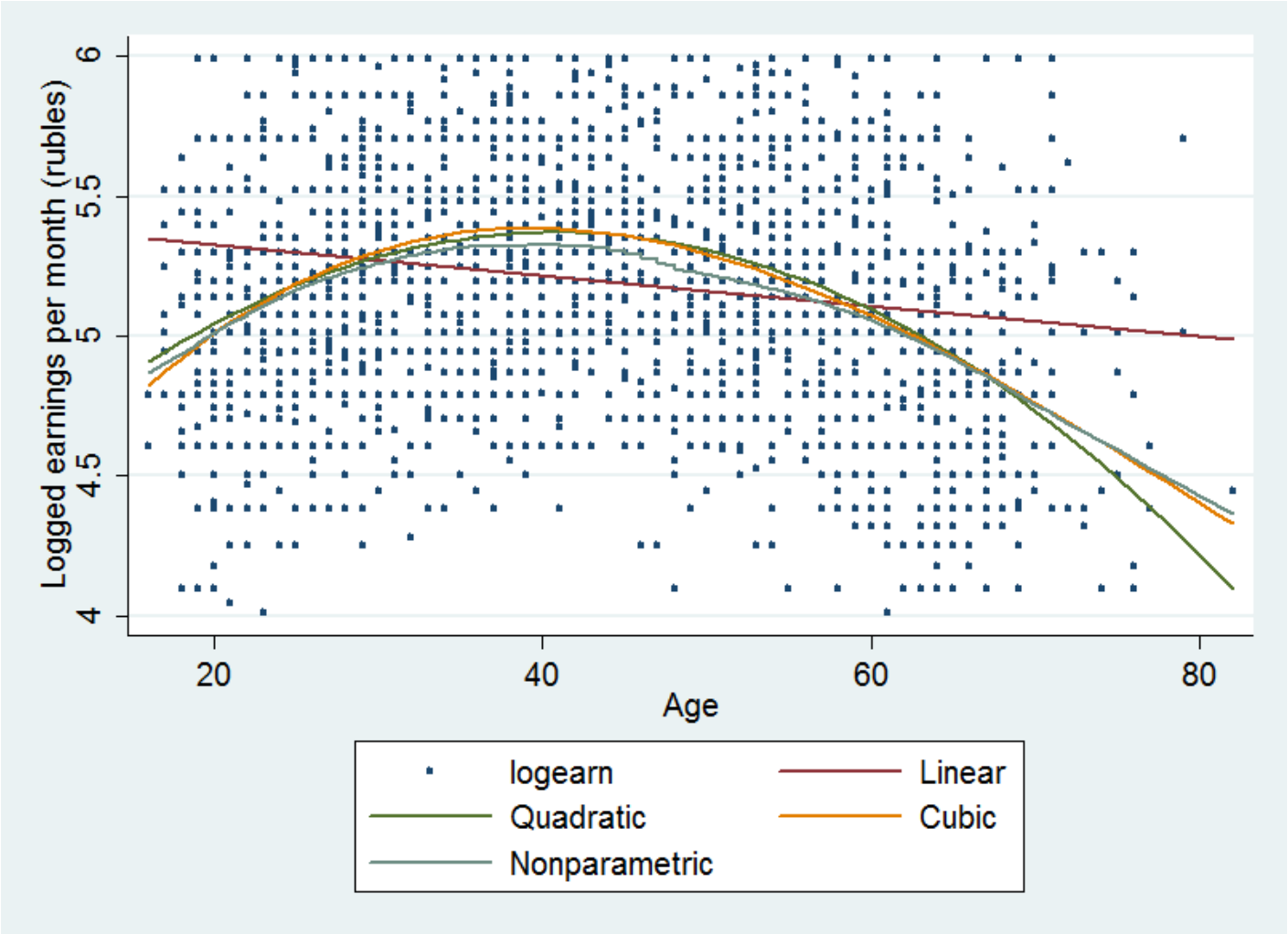
Высота потолка <- логарифм дохода

переменные	коэф.	ст.ошибка	t	p
логарифм дохода	3.35	1.18	2.8	< 0.05
константа	239	6	39	< 0.001
n	2,171			
R-квадрат	0.004			

Иногда требуется трансформировать зависимую переменную

- Средняя величина плохо описывает сильно смещенные распределения (напоминаю, что регрессия предсказывает условную среднюю величину)
- В этом случае часто используется логарифмическая трансформация – например, когда зависимой переменной является доход

Возраст и логарифм дохода



Логарифмическое преобразование

- $log 0$ не существует
- Чтобы решить эту проблему, мы можем добавить к доходу небольшую величину (например, 0.5)
- В качестве альтернативы, можно взять квадратный корень дохода

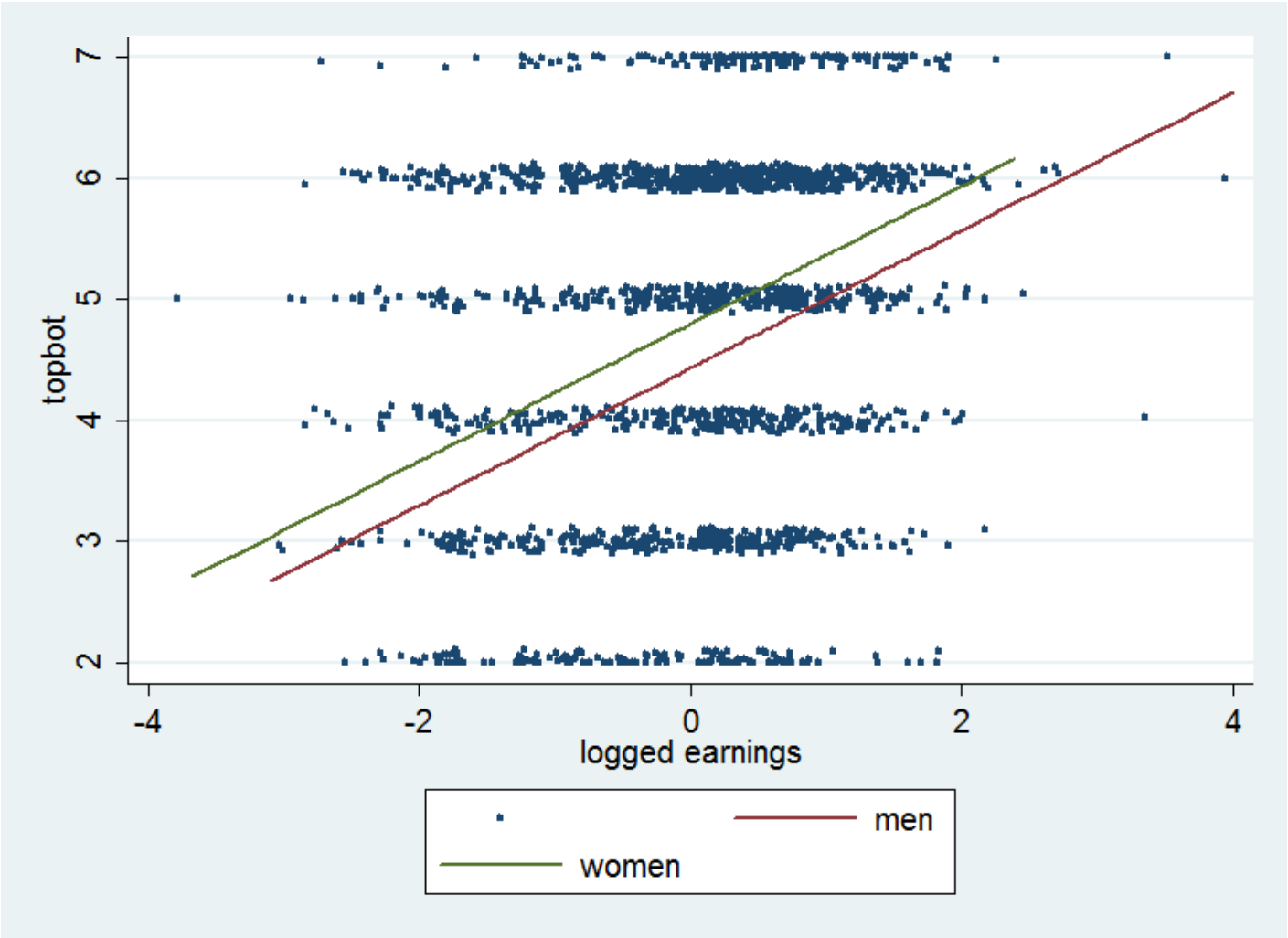
Взаимодействие между переменными

- В обычной множественной регрессии без эффектов взаимодействия мы оцениваем *средний* эффект переменной, контролируя по другим переменным
- Например, мы оцениваем связь образования и дохода, контролируя возраст. Мы моделируем среднюю связь образования и дохода для всех возрастных групп
- Связь образования и дохода может существенно различаться в возрастных группах
- В этом случае, необходимо использовать эффекты взаимодействия между переменными

Самооценка социального положения <- логарифм дохода и пол (ISSP Китай 2009)

переменные	коэф.	ст.ошибка	p
логарифм дохода (ст.)	0.57	0.04	< 0.001
мужчины	-0.37	0.08	< 0.001
константа	4.8	0.06	< 0.001
n	2434		
R-квадрат	0.08		

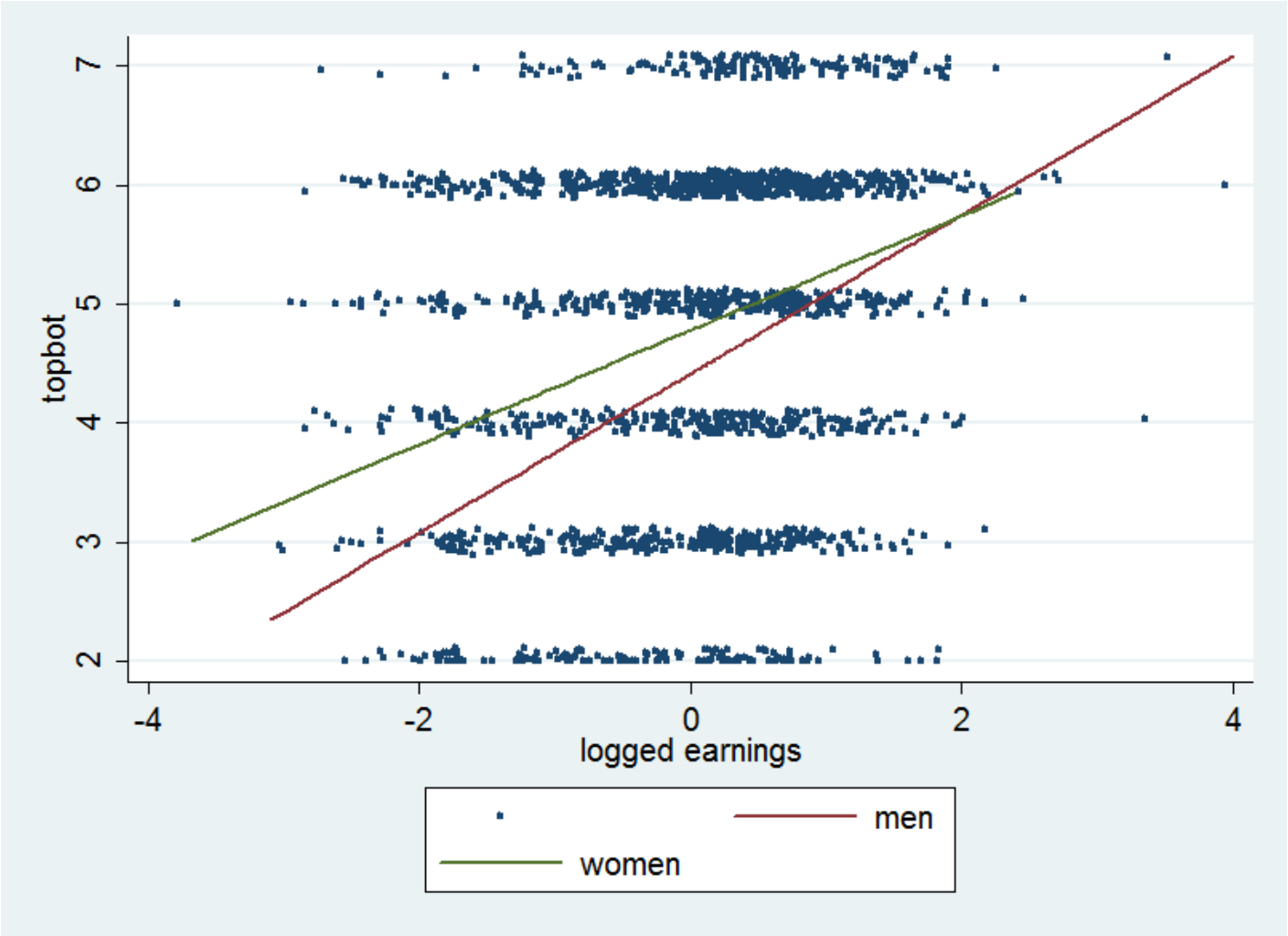
Связь дохода и самооценки социального положения для мужчин и женщин



Интепретация эффектов

- Линии параллельны, таким образом эффект дохода моделируется как одинаковый для мужчин и женщин
- Это может быть не так

Отдельные регрессии для мужчин и женщин



Взаимодействие между дихотомической и интервальной переменной

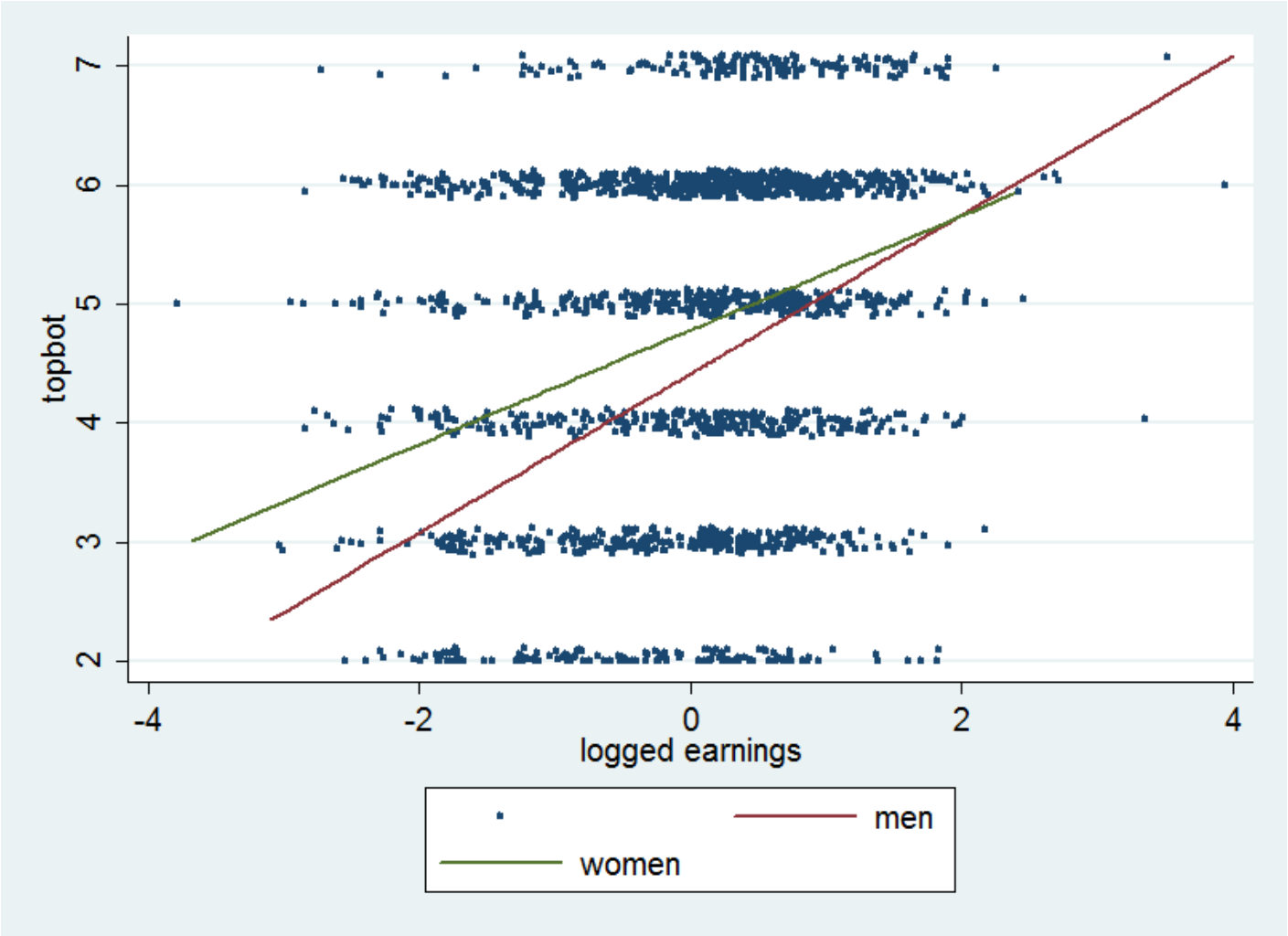
- Необходимо перемножить две переменные и добавить их произведение в регрессионное уравнение (оставив сами переменные)

Регрессия с эффектом взаимодействия

переменные	коэф.	ст.ошибка	p
логарифм дохода (ст.)	0.48	0.05	< 0.001
мужчины	-0.37	0.08	< 0.001
логарифм дохода * мужчины	0.19	0.08	0.02
константа	4.8	0.06	< 0.001
n	2434		
R-квадрат	0.08		

- $\text{topbot} = 4.8 + 0.48 * \text{logearn} - 0.37 * \text{male} + 0.19 * \text{logearn} * \text{male}$

Визуализация эффектов

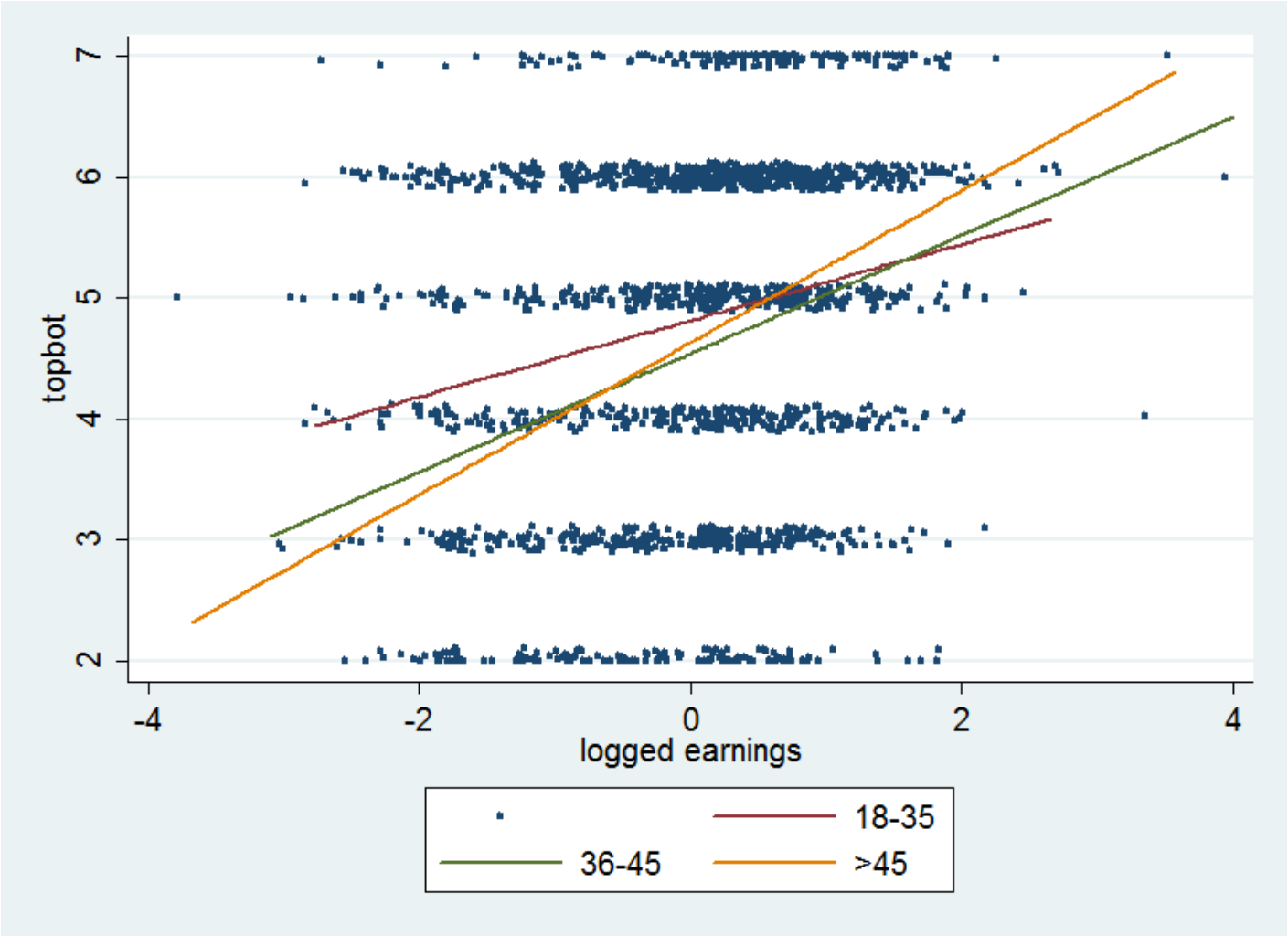


Взаимодействие между интервальной переменной и категориальной переменной с тремя и более категориями: доход и возрастная группа

переменные	коэф.	ст.ошибка	p
логарифм дохода (ст.)	0.32	0.10	< 0.01
возраст (баз. 18-30)			
$31 - 45$	-0.27	0.11	0.02
> 45	-0.18	0.11	0.11
логарифм дохода * 31-45	0.17	0.12	0.14
логарифм дохода * > 45	0.31	0.12	< 0.01
константа	4.8	0.10	< 0.001
n	2434		
R-квадрат	0.08		

- $\text{topbot} = 4.8 + 0.32 * \text{logearn} - 0.27 * \text{midage} - 0.18 * \text{mature} + 0.17 * \text{midage} * \text{logearn} + 0.31 * \text{mature} * \text{logearn}$

Визуализация эффектов



Взаимодействие двух интервальных переменных: доход и образование как предикторы самооценки социального положения

переменные	коэф.	ст.ошибка	p
логарифм дохода (ст.)	0.43	0.05	< 0.001
образование (годы, ст.)	0.21	0.05	< 0.001
логарифм дохода * образование	0.02	0.04	0.57
константа	4.6	0.04	< 0.001
n	2412		
R-квадрат	0.08		

Взаимодействие двух категориальных переменных

- Необходимо создать переменную для эффекта взаимодействия для каждой комбинации фиктивных переменных

Взаимодействие двух категориальных предикторов: возрастная группа и пол

переменные	коэф.	ст.ошибка	p
мужчины	-0.35	0.15	0.02
возраст (баз. 18-30)			
$31-45$	-0.49	0.13	<0.001
>45	-0.71	0.13	<0.001
мужчины * 31-45	0.17	0.19	0.37
мужчины * >45	0.44	0.19	0.02
константа	5.1	0.11	<0.001
n	3010		
R-квадрат	0.01		

Регрессионное уравнение

- $\text{topbot} = 5.1 - 0.35 * \text{male} - 0.49 * \text{midage} - 0.71 * \text{mature} + 0.17 * \text{midage} * \text{male} + 0.44 * \text{mature} * \text{male}$
  - Молодые женщины: $\text{topbot} = 5.1$
  - Женщины старше 45: $\text{topbot} = 5.1 - 0.71 = 4.39$
  - Молодые мужчины: $\text{topbot} = 5.1 - 0.35 = 4.75$
  - Мужчины старше 45: $\text{topbot} = 5.1 - 0.35 - 0.71 + 0.44 = 4.48$
- Эффект возраста для женщин сильнее, чем эффект возраста для мужчин

Взаимодействие трех переменных

- Возможно взаимодействие трех (и более) переменных, но оно редко встречается
- Пример: есть ли отличия в разнице в силе эффекта возраста между мужчинами и женщинами в разных странах? (Взаимодействие пола, возраста и страны.)