Регрессия: допущения, диагностика и некоторые советы

А.Р.Бессуднов
(bessudnov@gmail.com)

НИУ ВШЭ, факультет социологии

16 марта 2012

Прошлая лекция

- Статистический вывод (statistical inference) в регрессии
- Преобразование переменных в регрессионных моделях и нелинейность
- Эффекты взаимодействия

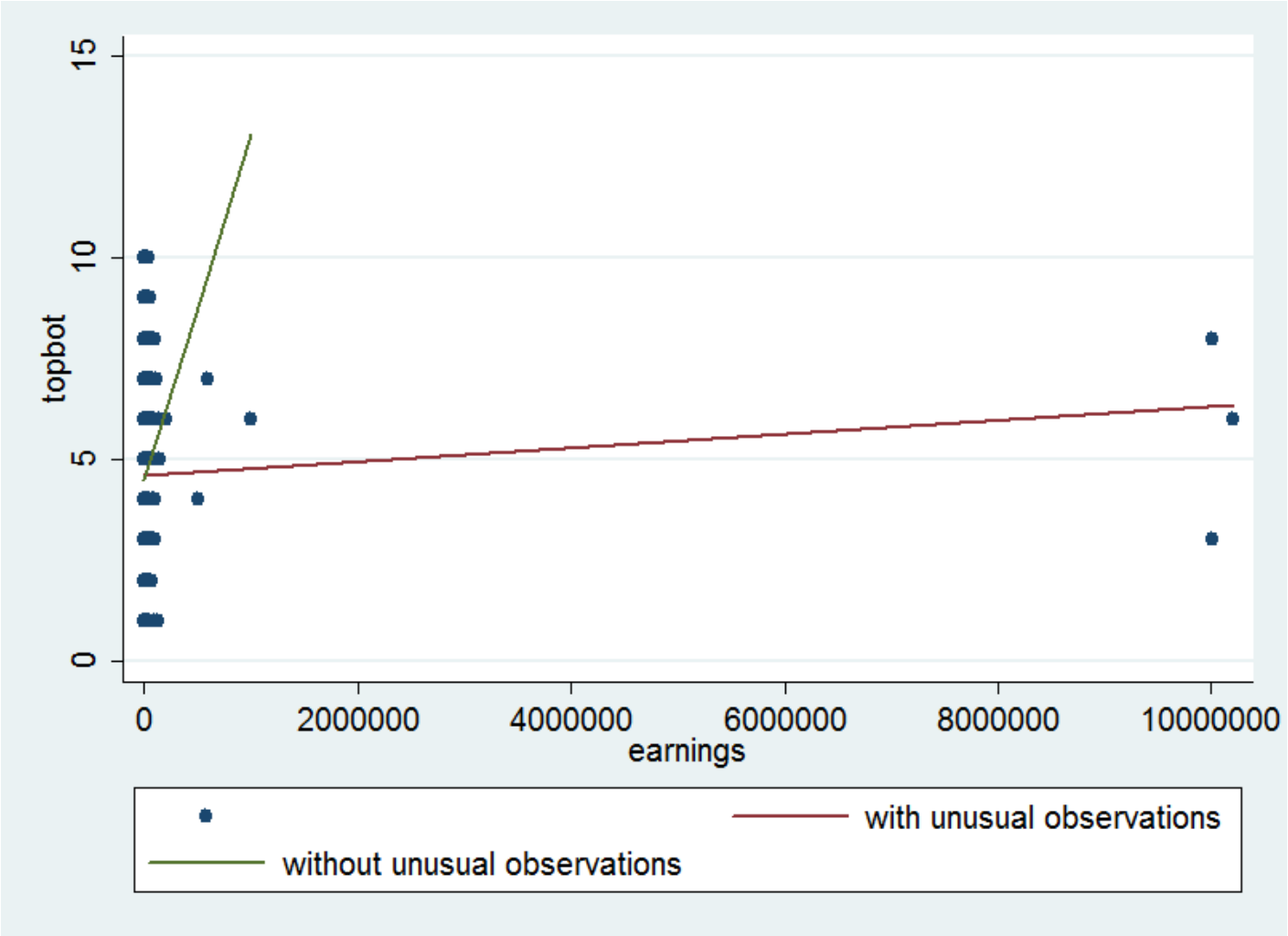
Сегодня

- Допущения регрессионной модели и диагностика
- Несколько советов, касающихся регрессионного моделирования

Статистические “выбросы” (outliers)

- Необычные наблюдения могут сильно влиять на регрессионные коэффициенты (и другие статистические параметры)
- В контексте множественной регрессии, “выброс” – это наблюдение, которое принимает необычное (слишком большое или слишком малое значение), учитывая значения предикторов

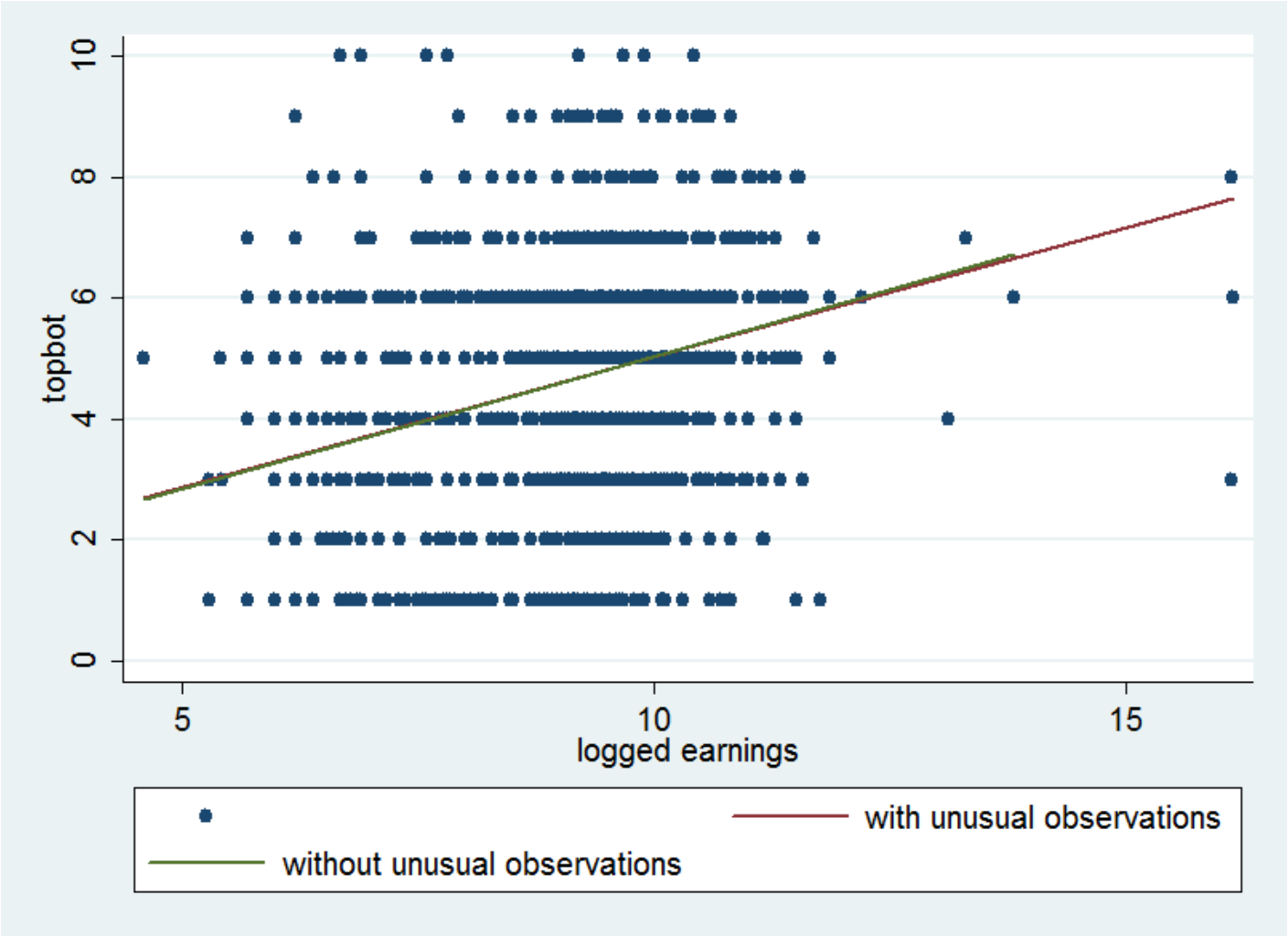
Две регрессии: с необычными наблюдениями и без них



Решение

- В данном случае одним из решений будет взять логарифм дохода (или просто удалить эти наблюдения)

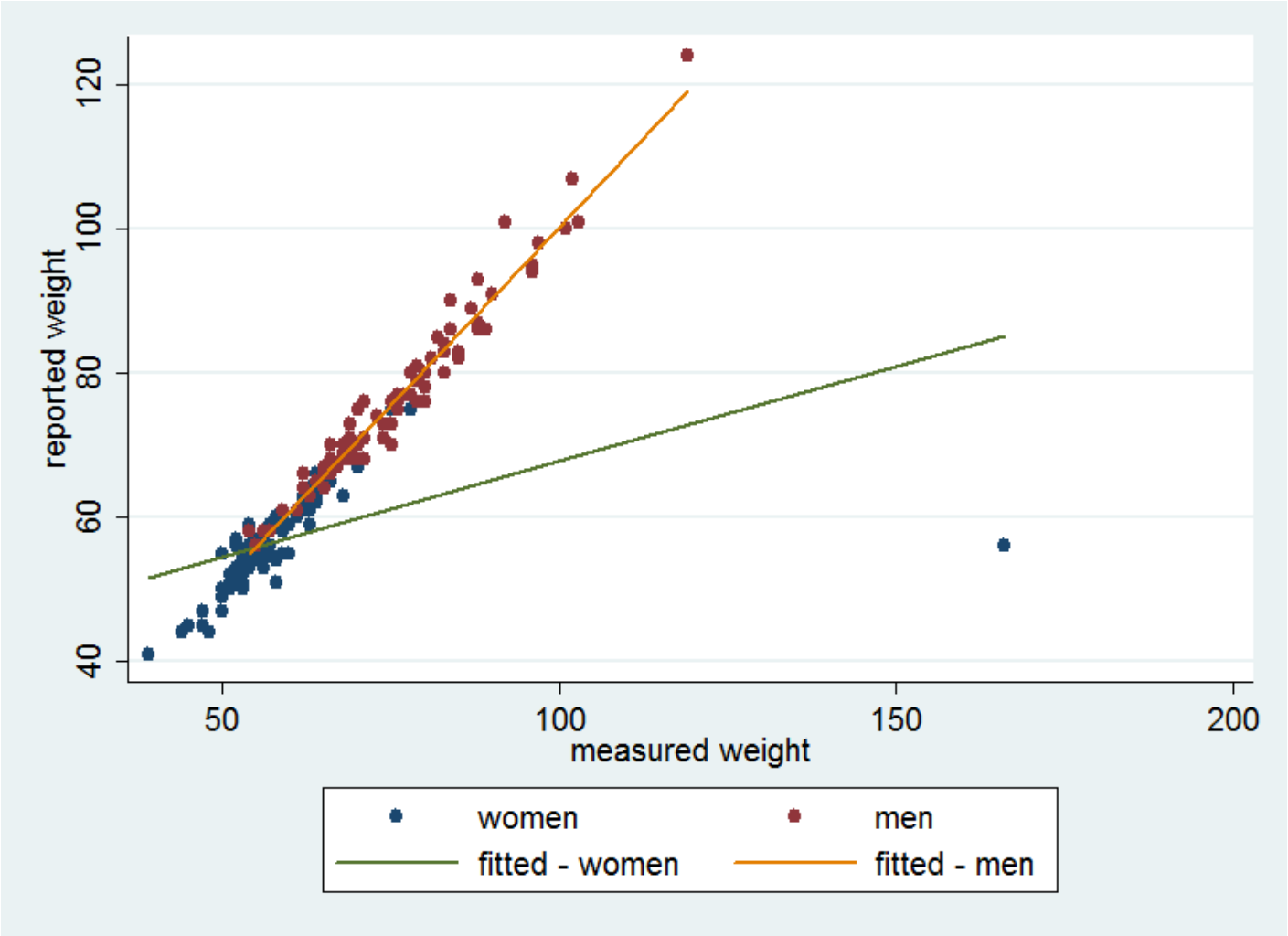
topbot и логарифм дохода



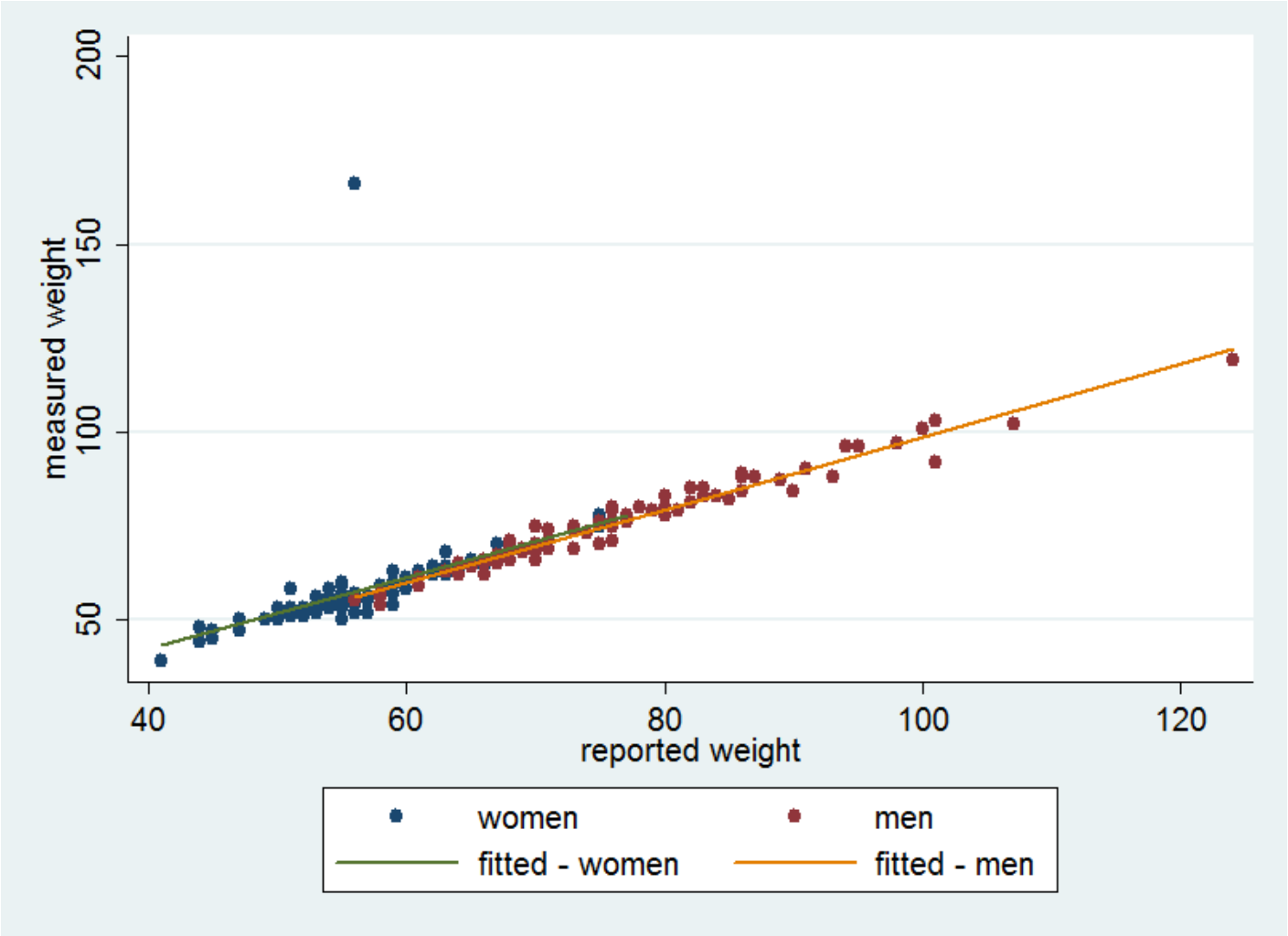
Формальная диагностика

- Существуют более формальные методы для идентификации выбросов
- Разница между “несогласием” (discrepancy) и “рычагом” (leverage)
  - Несогласие: насколько велики остатки (насколько далеко реальное наблюдение находится от регрессионнойй линии или плоскости)
  - “Рычаг”: в какой части распределения независимой переменной находится наблюдение
- Влияние на регрессионные коэффициенты определяется одновременно несогласием и “рычагом”

Пример (данные Дэвиса)



Пример (2)


measured weight
200
150
100
50
40
60
80
100
120
reported weight
women
men
fitted - women
fitted - men

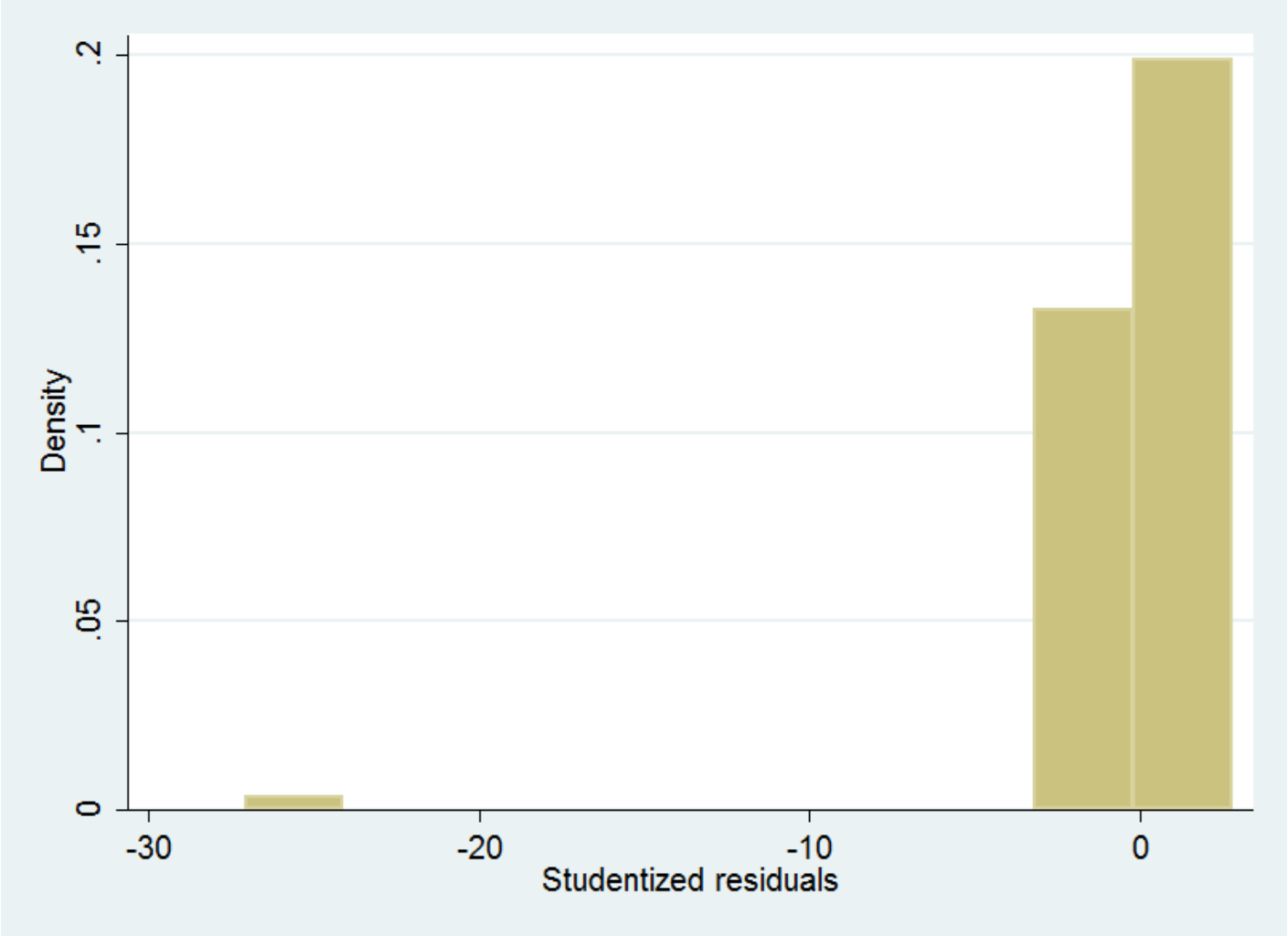
Показатель “рычага”: hat values

- Hat values показывают, насколько далеко наблюдения лежат от средней x (в парной регрессии)
- $h_i = \frac{1}{n} + \frac{(X_i - \bar{X})^2}{\sum_{j=1}^{n}(X_j - \bar{X})^2}$
- “Рычаг” не зависит от зависимой переменной

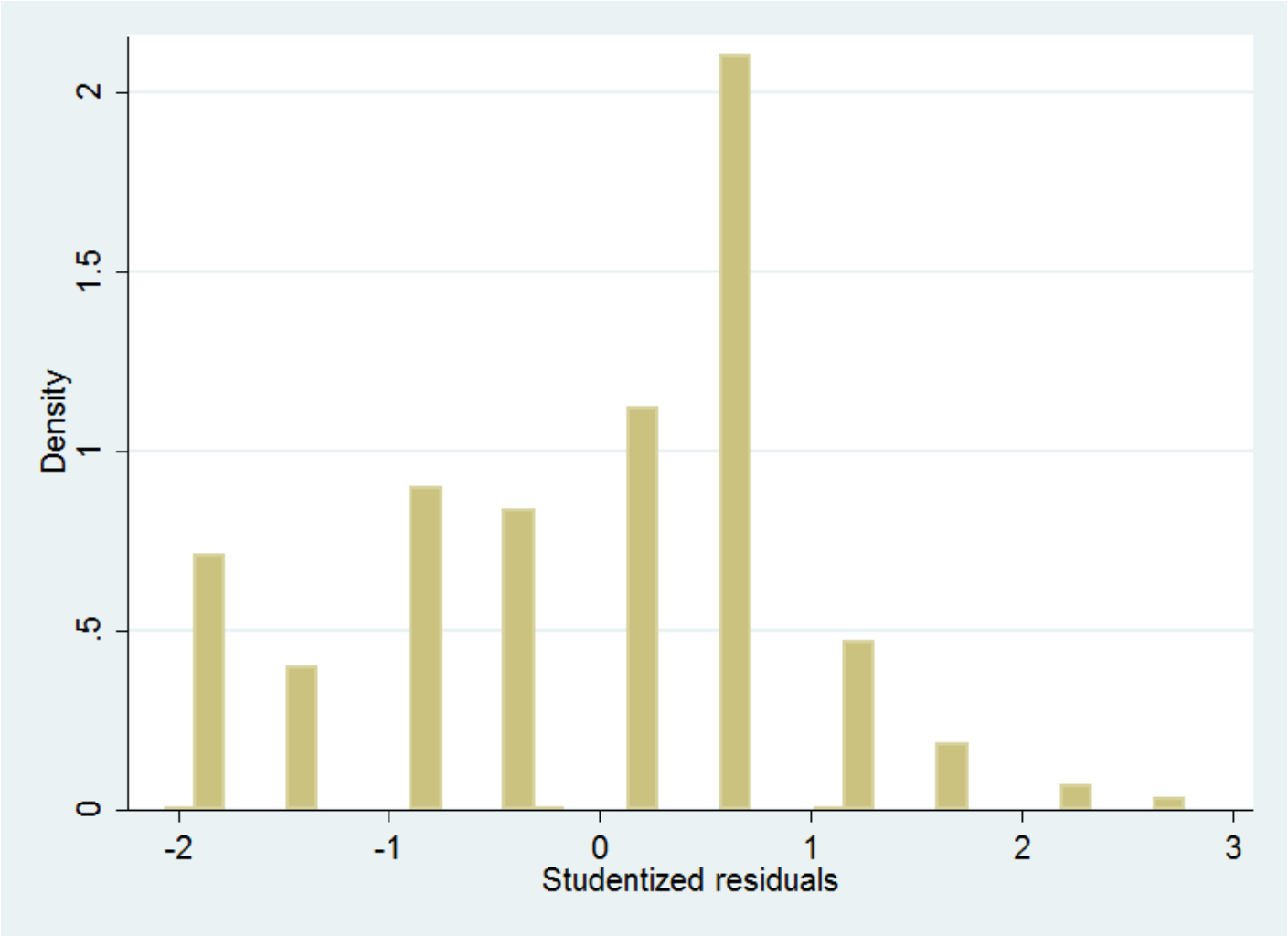
Показатель “несогласия”: Стьдентизированные остатки

- Регрессионные остатки показывают, насколько “необычны” наблюдения в регрессии
- Стьюдентизированные остатки: $E_i^* = \frac{E_i}{S_{E(-i)}\sqrt{1-h_i}}$

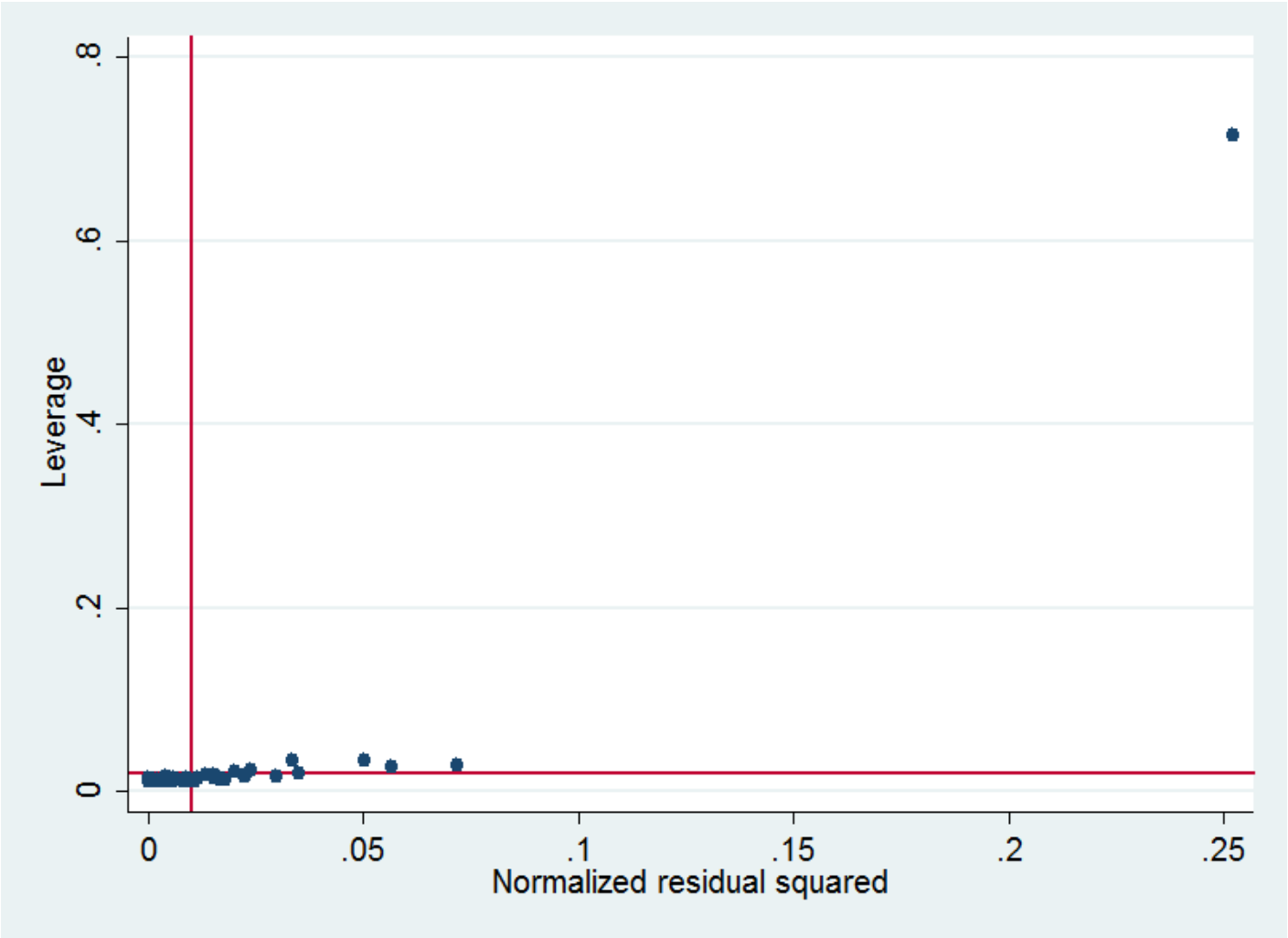
Гистограмма стьюдентизированных остатков для данных Дэвиса (женщины, rep weight ← weight)



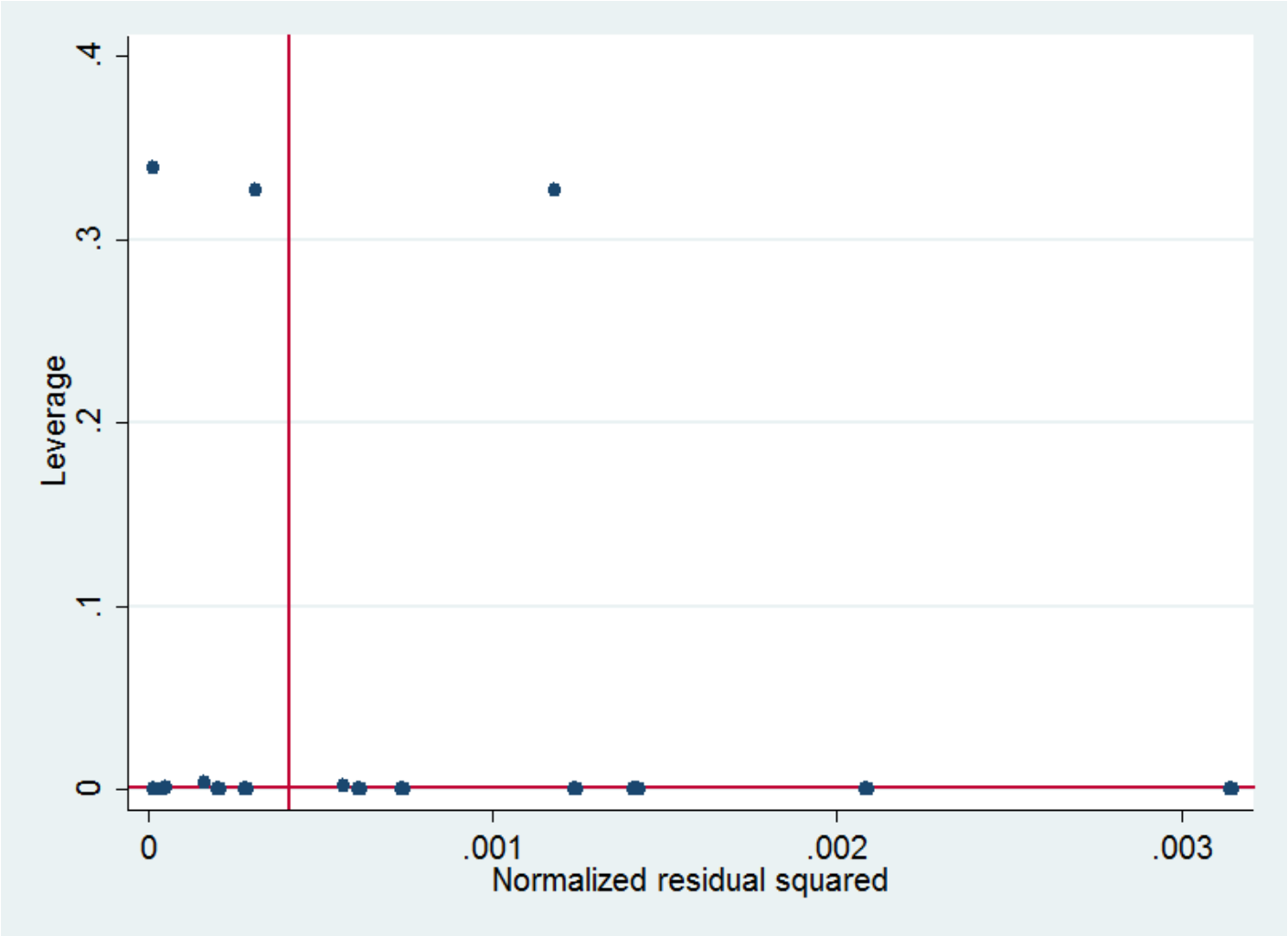
Для данных о доходе в Китае



Стьюдентизированные остатки и hat values: данные Дэвиса



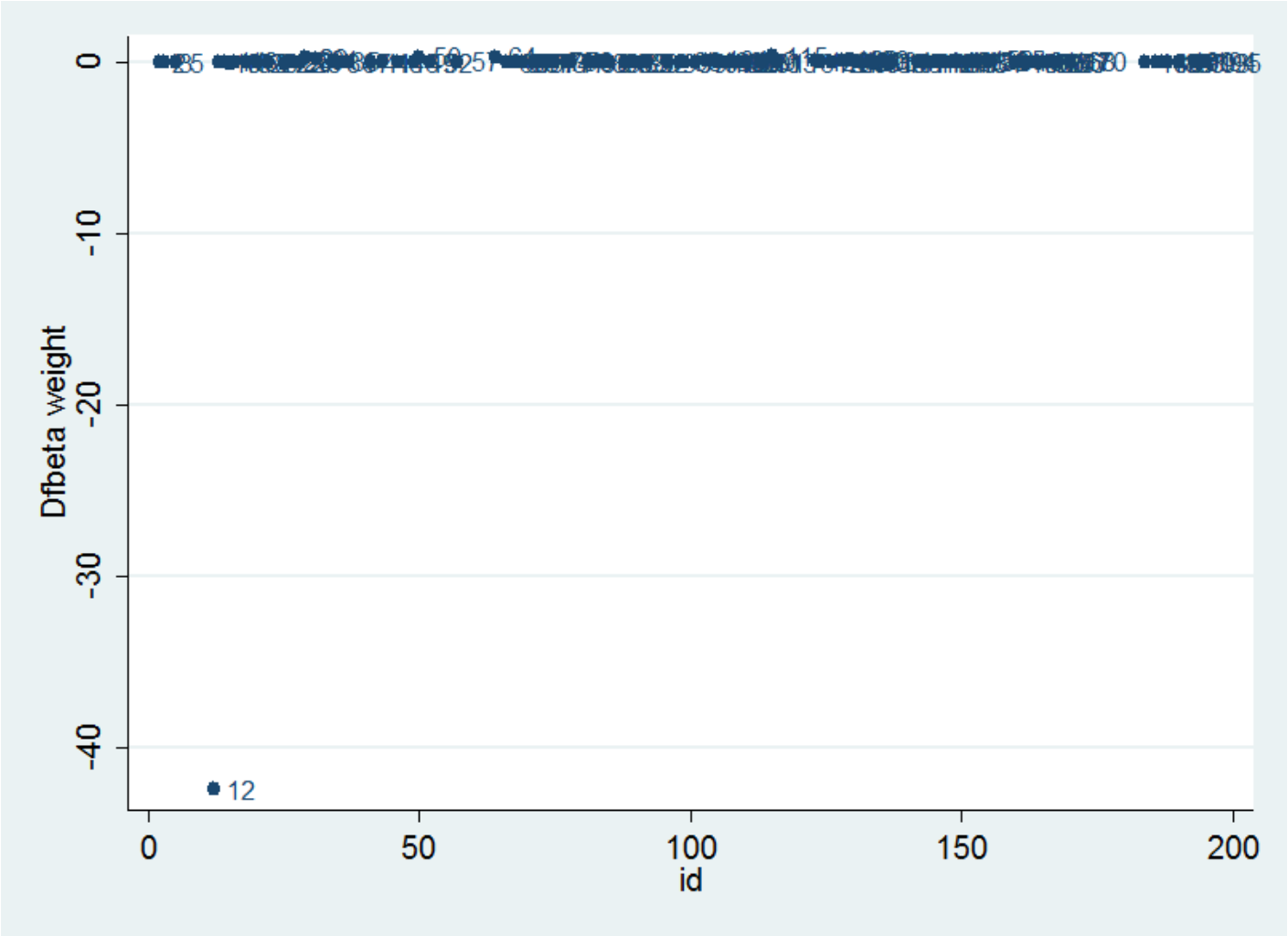
Данные о доходе в Китае



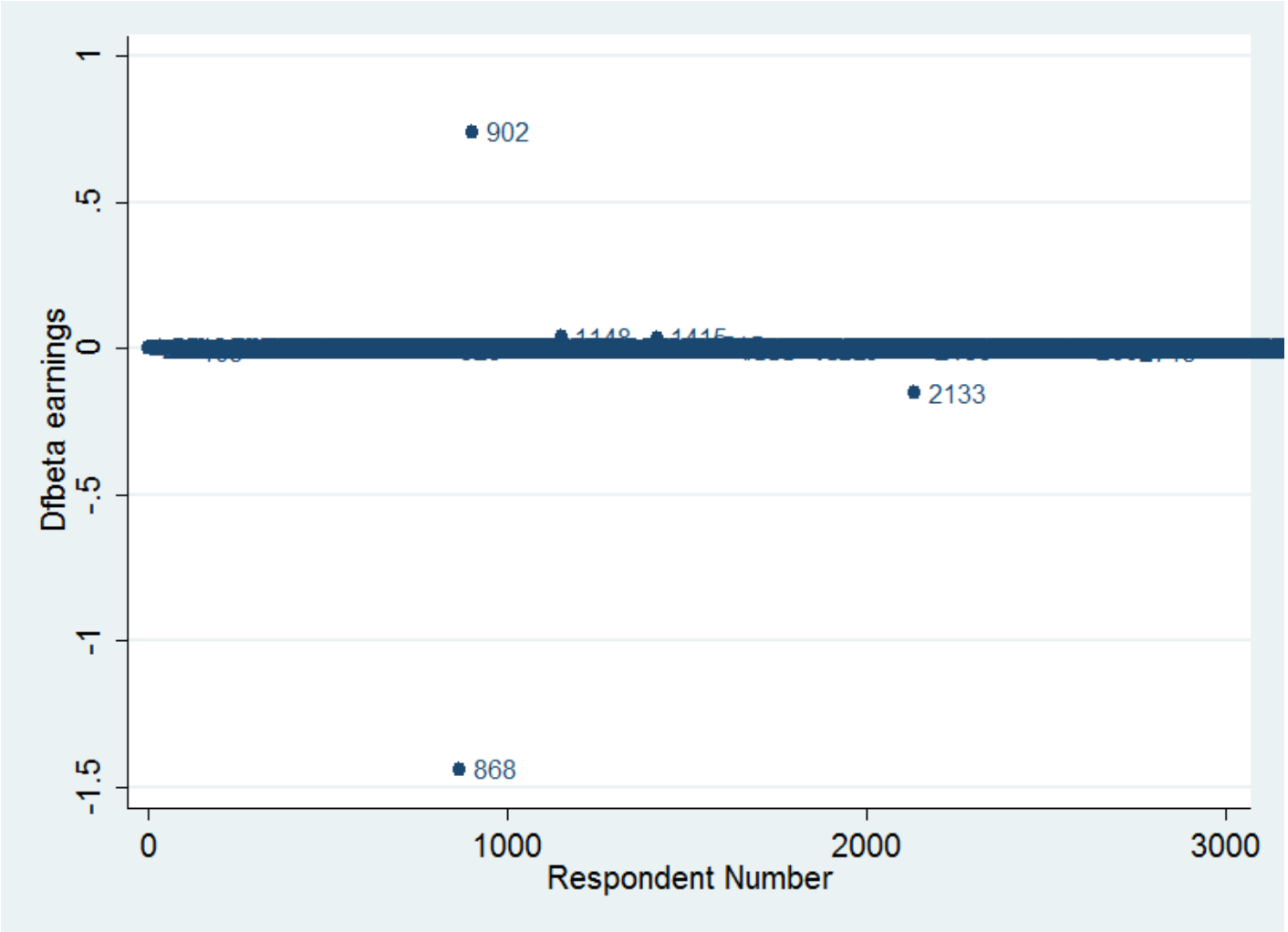
Влияние на регрессионные коэффициенты

- Влияние на коэффициенты определяется совместно “несогласием” и “рычагом”
- Статистика, которая измеряет влияние, называется DFBETA. Эта статистика показывает, как изменится регрессионный коэффициент (в стандартных отклонениях), если наблюдение удалить из базы данных

DFBETAS для данных Дэвиса



DFBETAS для китайских данных

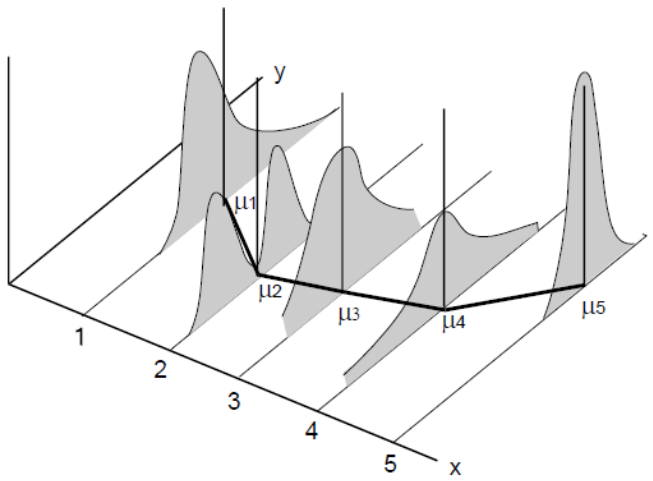


Что делать с выбросами

- Необходимо понять причину возникновения “выбросов”
- Если это ошибка кодирования, можно перекодировать или удалить наблюдения
- Если это реальное наблюдение, то удалять такие наблюдения нужно с осторожностью. Можно преобразовать переменную или добавить в модель другие переменные, которые позволят “объяснить” выбросы
- В больших выборках редко обнаруживаются выбросы. Соответственно, для политологов выбросы – бóльшая проблема, чем для социологов

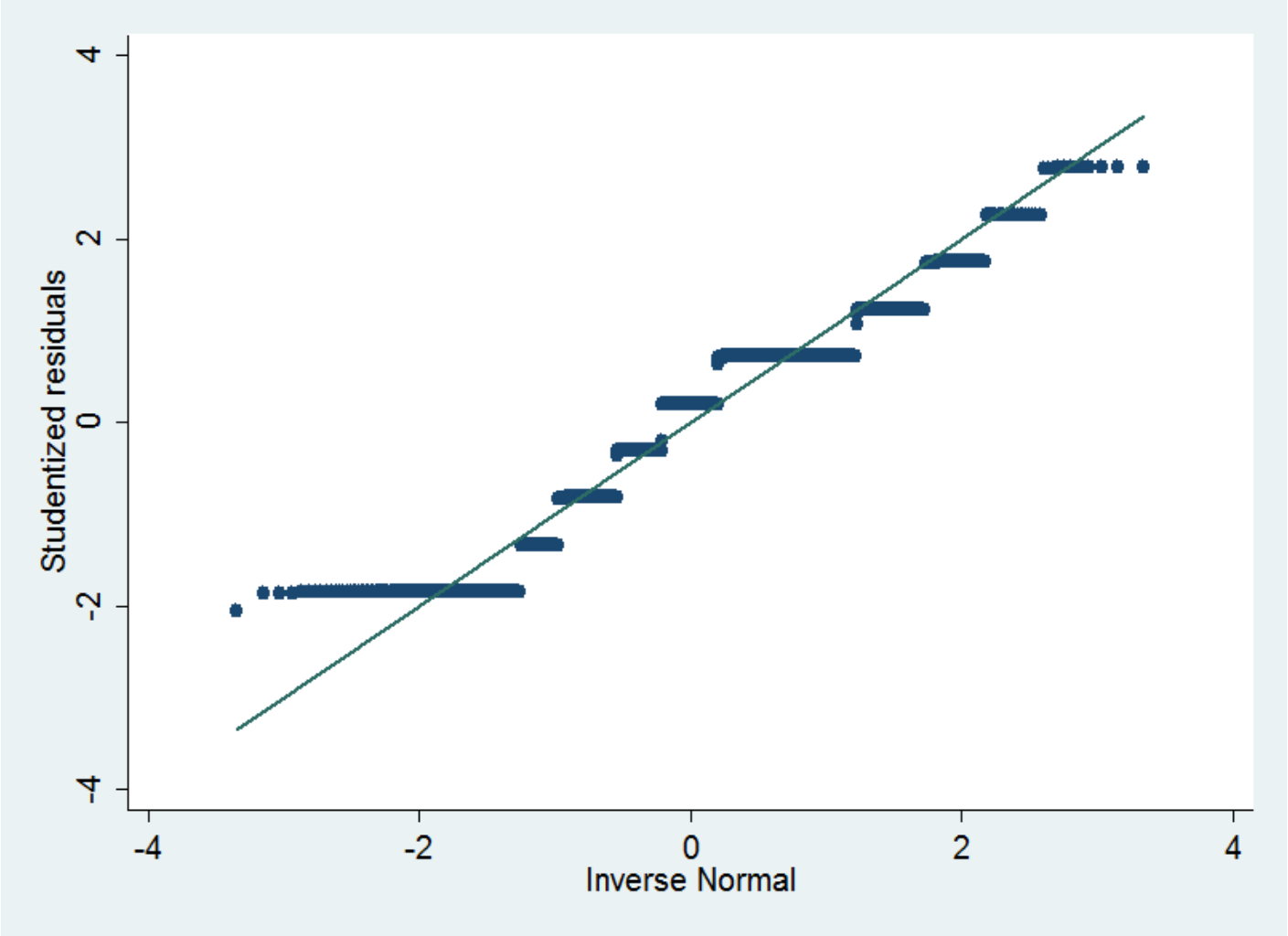
Отличное от нормального распределение остатков

- Одним из допущений линейной регрессии является то, что остатки в генеральной совокупности нормально распределены
- Мы можем проверить истинность этого допущения в выборке



(Источник: Дж.Фокс)

Нормальный вероятностный график (Q-Q plot) для китайских данных



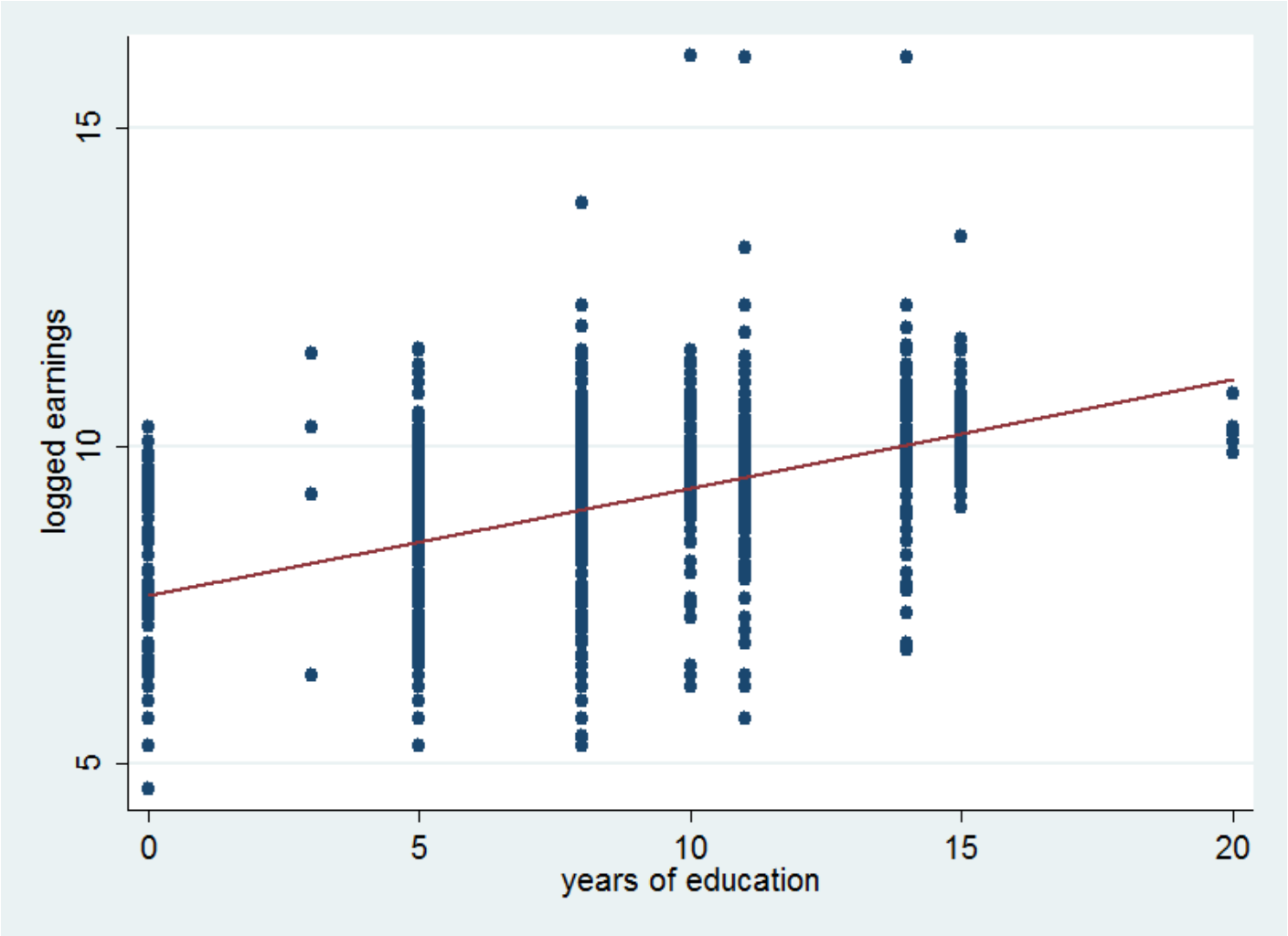
Распределение остатков не является нормальным

- Это не влияет на регрессионные коэффициенты, но влияет на стандартные ошибки
- Одним из возможных решений является преобразование зависимой переменной
- Другим решением является робастная регрессия

Непостоянство дисперсии остатков (гетероскедастичность)

- Другим допущением МНК-регрессии является постоянство дисперсии остатков на разных уровнях x
- Это допущение называется гомоскедастичностью, если оно нарушается – гетероскедастичность

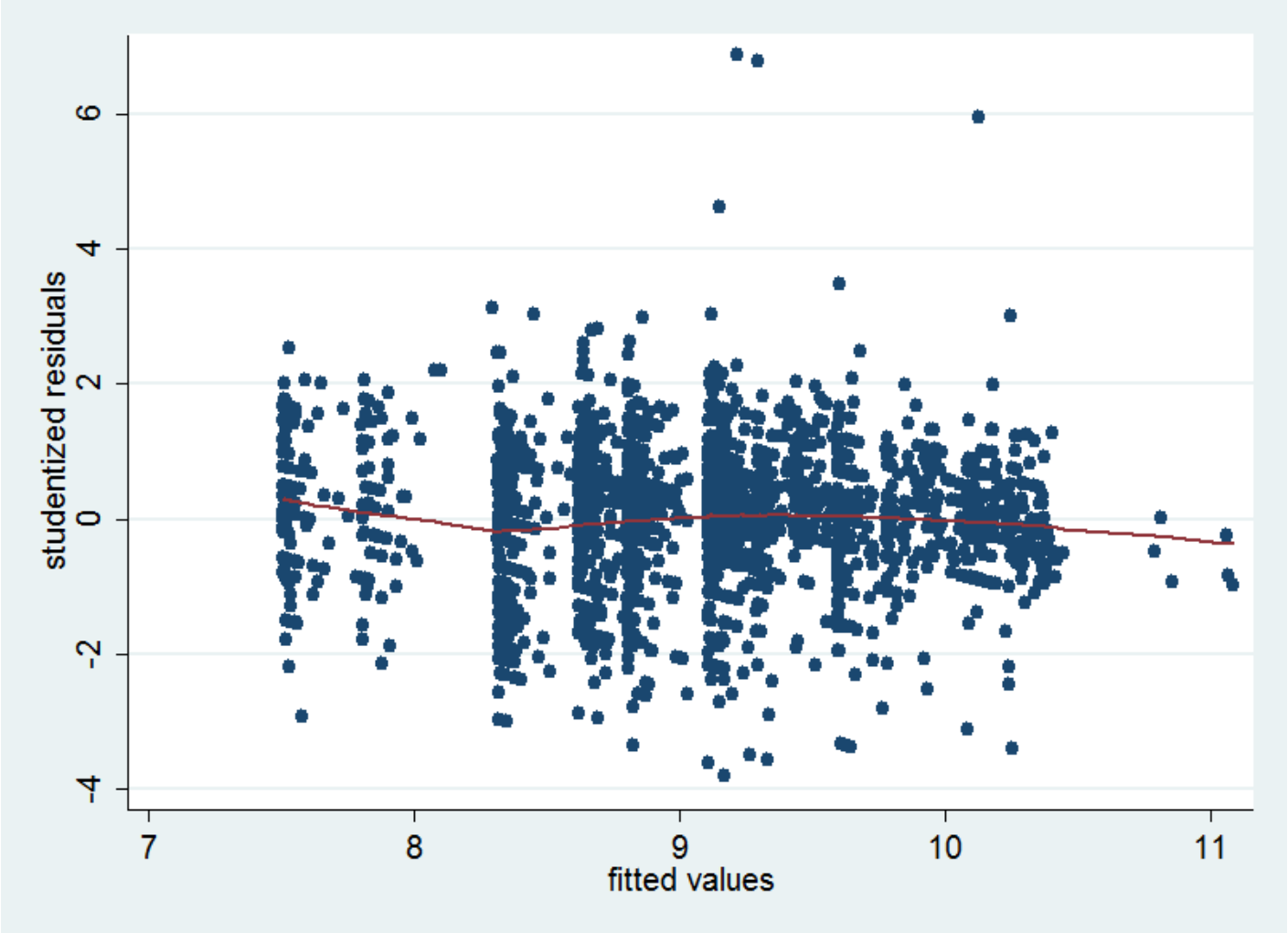
Логарифм дохода <- образование



Множественная регрессия

- Для множественной регрессии мы можем посмотреть на диаграмму рассеивания остатков и предсказанных значений
- Если между предсказанными значениями и остатками есть зависимость, то это указывает на гетероскедастичность

Логарифм дохода <- образование, возраст, возраст в квадрате, пол



Решение проблемы гетероскедастичности

- Гетероскедастичность не влияет на регрессионные коэффициенты, но влияет на стандартные ошибки
- Это влияние является серьезным только в том случае, если максимальная дисперсия остатков превышает минимальную в 10 раз (согласно более консервативному подходу, в 4 раза)
- Возможные решение: преобразование зависимой переменной, использование метода взвешенных наименьших квадратов (WLS), робастная регрессия

Нелинейность

- Допущением линейной регрессии является то, что связь между зависимой и независимой переменной наилучшим образом описывается линейной функцией
- Часто это не так: мы подробно обсуждали это на прошлом занятии
- Решения: преобразования независимых переменных, непараметрическая регрессия

Мультиколлинеарность

- Мультиколлинеарностью называется ситуация. когда две или более независимые переменные сильно скоррелированы между собой ($r > 0.8$)
- В этом случае тяжело одновременно оценить эффекты этих переменных
- Статистика, которая используется для идентификации мультиколлинеарности, называется VIF (variance-inflation factor)
- Для переменной x_j, $VIF_j = \frac{1}{1-R_j^2}$, где R_j^2 – R^2 для регрессии x_j на другие предикторы в модели
- VIF измеряет влияние мультиколлинеарности на точность оценки b_j
- Если $VIF > 10$, возможна мультиколлинеарность
- Если вы одновременно используете переменную и квадрат этой переменной в модели, на VIF для этих переменных обращать внимание не следует

Решения проблемы мультиколлинеарности

- Редко встречается в социологии (чаще – в политологии)
- Возможные решения: удаление из модели одной из переменных, создание шкалы с использованием факторного анализа

Некоторые советы, касающиеся регрессионного анализа

- Обращайте внимание на разницу между статистической и практической значимостью

Корреляция не означает причинно-следственную связь

- Регрессия не идентифицирует причинно-следственные связи: 1) обратная причинно-следственная зависимость, 2) ненаблюдаемые факторы
- Корректнее говорить о статистической связи, а не о влиянии

Начинайте с простого анализа

- “I find myself telling people to go simple, simple, simple. When someone gives me their regression coefficient I ask for the average, when someone gives me the average I ask for a scatterplot, when someone gives me a scatterplot I ask them to carefully describe one data point, please” (A.Gelman)

Не добавляйте в модель слишком много независимых переменных

- В множественной регрессии вы интерпретируете регрессионные коэффициенты, удерживая другие переменные в модели на одной уровне. В некоторых случаях это не имеет смысла

Используйте графики для иллюстрации моделей

- Графики чаще позволяют лучше отобразить данные и модель, чем таблицы (однако таблицами также не следует пренебрегать)

Следующий модуль

- Регрессионные модели для дихотомических зависимых переменных
- Регрессионные модели для категориальных зависимых переменных
- Логарифмически-линейные модели для таблиц сопряженности
- Введение в причинно-следственный анализ