Линейная регрессия и корреляция

А.Р.Бессуднов
(bessudnov@gmail.com)

НИУ ВШЭ, факультет социологии

17 февраля 2012

Что такое регрессионный анализ?

- Регрессионный анализ – это широкий класс статистических методов, которые изучают как связана одна (зависимая) переменная и несколько других (независимых) переменных
- Например, нас может интересовать зависимость дохода от таких переменных, как пол, возраст, образование и место жительства
- Другой пример: зависимость вероятность поддержать на президентских выборах того или иного кандидата от частоты пользования Интернетом и наличия опыта путешествий за границу
- В зависимости от типа данных и переменных и исследовательских задач используются разные варианты регрессии: линейная регрессия, логит и пробит-модели, регрессия Пуассона, негативная биномиальная регрессия, регрессия Кокса и т.д.
- Мы (пока) будем говорить об обычной линейной регрессии, оцениваемой методом наименьших квадратов, однако большая часть того, что будет сказано, относится и к другим методам

Почему нужно знать регрессионный анализ?

- В трех выпусках журнала American Sociological Review (один из двух наиболее престижных журналов в социологии) за август-декабрь 2011 г. было опубликовано 20 статей. Из них в 15 использовались статистические методы, и все 15 основаны на той или иной форме регрессии (намного сложнее тех, которые мы будем разбирать в этом курсе)
- Регрессионный анализ сейчас является основной формой статистического анализа данных в социальных науках, и без этого знания быть академическим социологом невозможно
- В маркетинговых исследованиях регрессионный анализ применяется редко, и там более распространены такие методы, как факторный и кластерный анализ

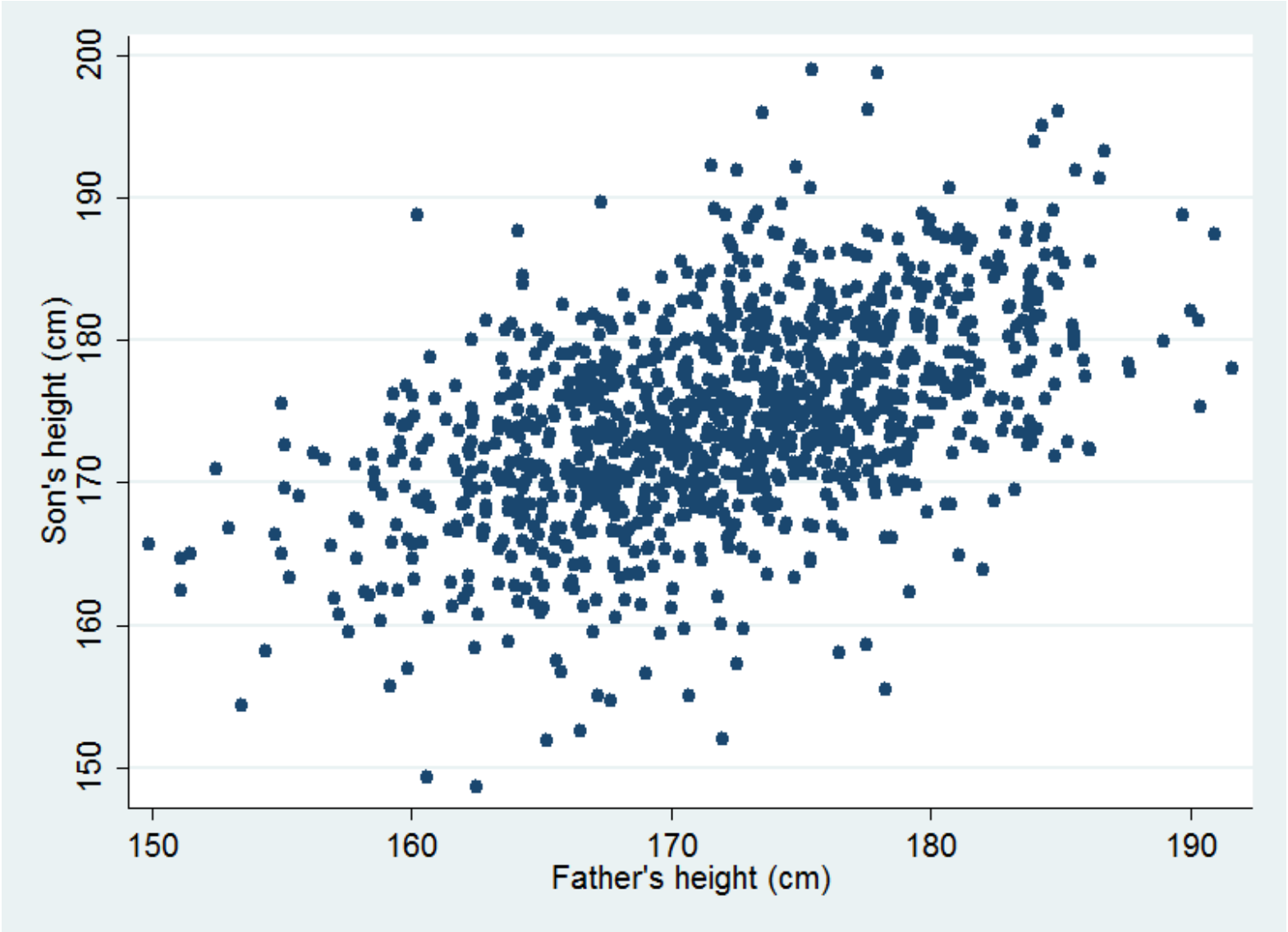
Краткая история регрессионного анализа

- Фрэнсис Галтон (1870-е гг.): английский биолог, двоюродный брат Дарвина, изучение наследственности
- Карл Пирсон (1880-90-е гг.): подвел математическое основание под корреляцию и регрессию
- 1960-70-е гг.: массовое применение регрессии (и основанном на ней метода структурных уравнений) в американской социологии (Питер Блау и Отис Дадли Данкан)
- Эконометрическая традиция

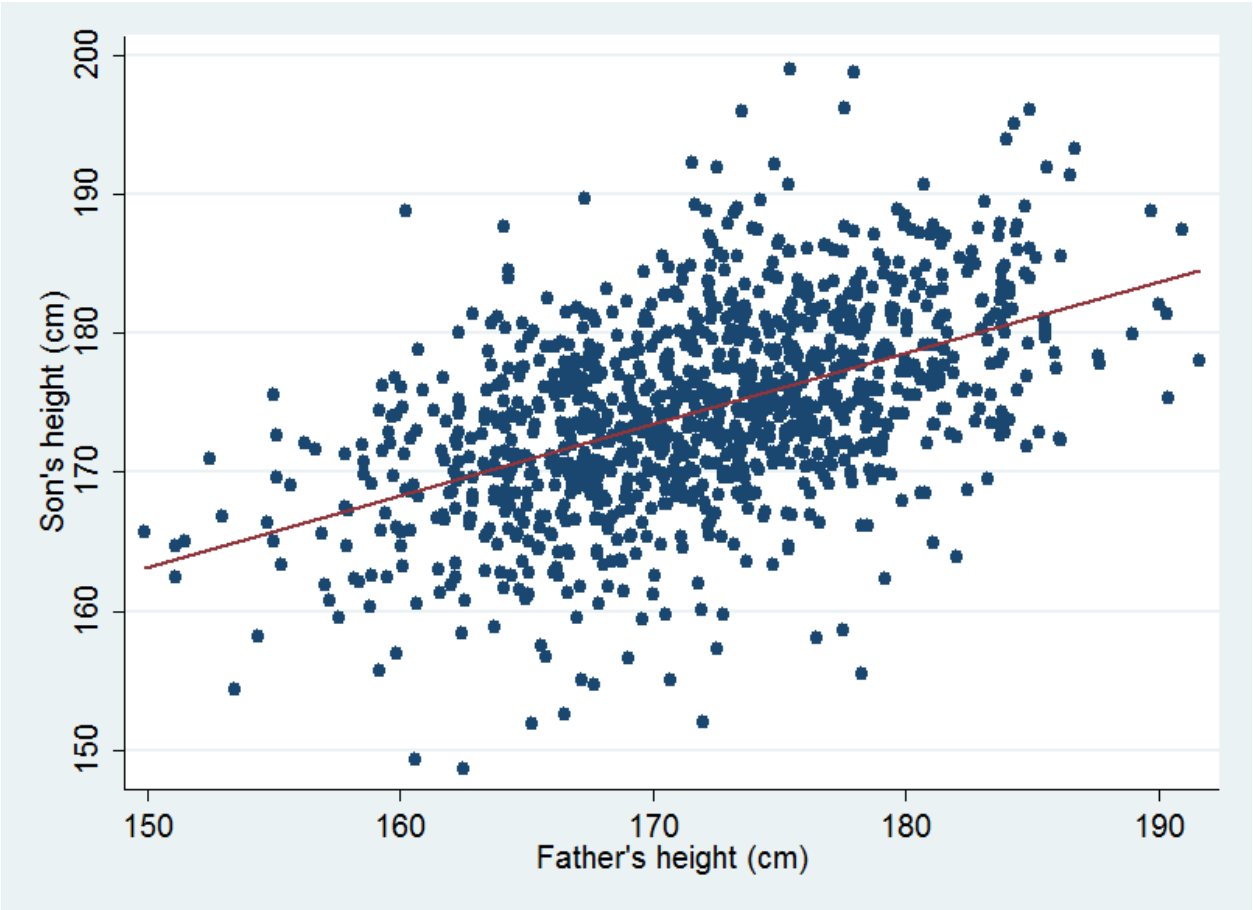
Когда используется линейная регрессия?

- Зависимая переменная должна быть интервальной или близкой к ней (например, доход, шкала удовлетворенности жизнью от 1 до 10, рост и т.д.)
- Независимые переменные могут быть любыми
- Мы начнем с простого случая, когда независимая переменная одна

Данные Галтона: рост отцов и рост сыновей



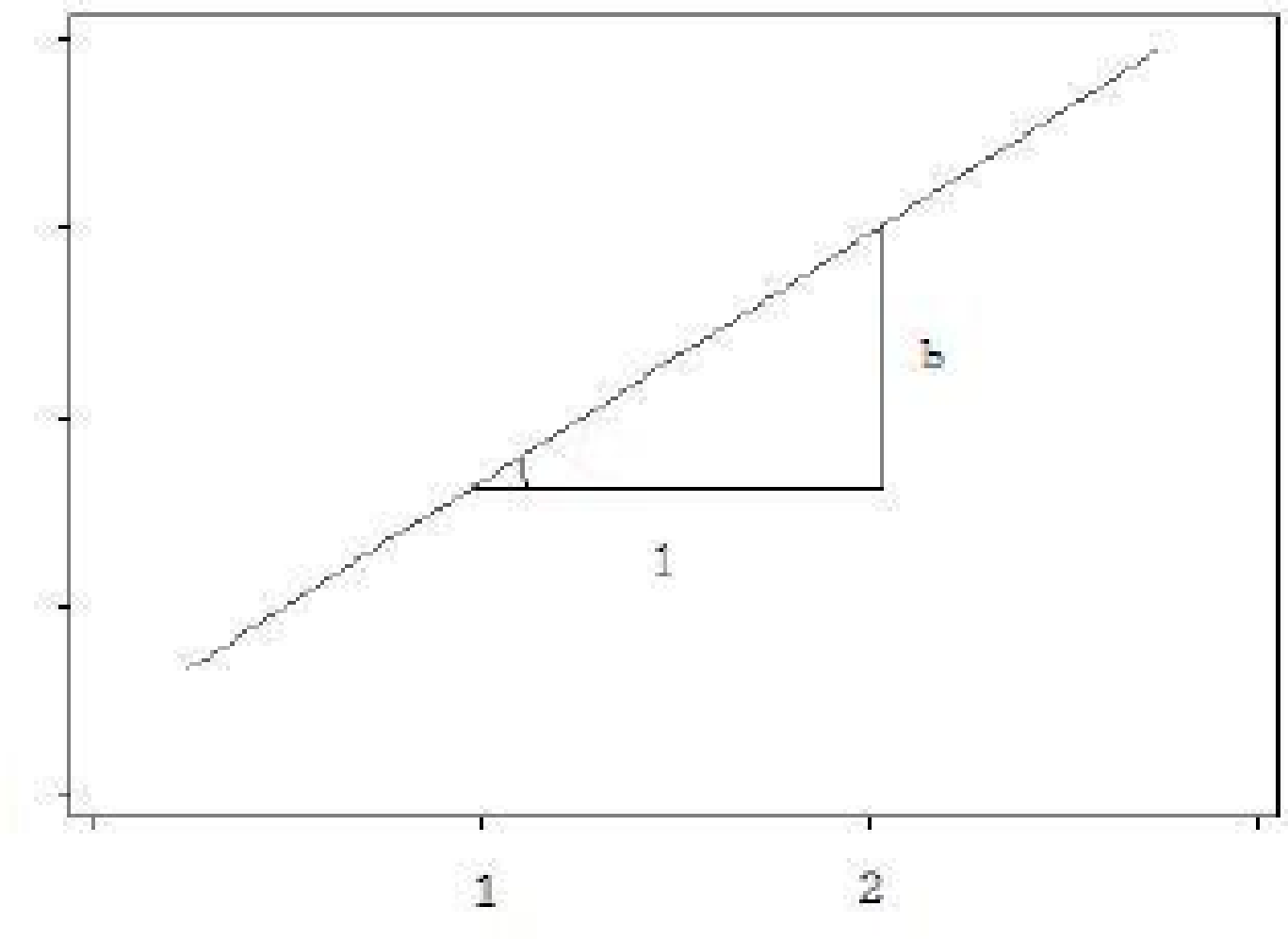
Наша задача заключается в том, что описать эти данные с помощью прямой линии. Это называется линейной регрессией



Регрессионное уравнение

- Прямая линия описывается формулой $y = a + bx$, в этом случае: $\mathrm{son'sheight} = a + b * \mathrm{father'sheight}$
- a – константа, “двигает” регрессионную линию вверх и вниз
- b – регрессионный коэффициент, меняет угол наклона линии. Математически, b – это тангенс угла, образуемого регрессионной линией и осью x

Геометрический смысл



Интерпретация регрессионного коэффициента b

- b показывает, на сколько единиц изменится y, если x меняется на 1
- В случае с нашем примером, b показывает, на сколько сантиметров в среднем отличается рост сыновей, чьи отцы отличались в росте на один сантиметр
- b нельзя интерпретировать как средний эффект изменения x для единицы анализа (человека, страны). b не измеряется причинно-следственную связь, а просто констатирует разницу в средних в группах

Интерпретация b (2)

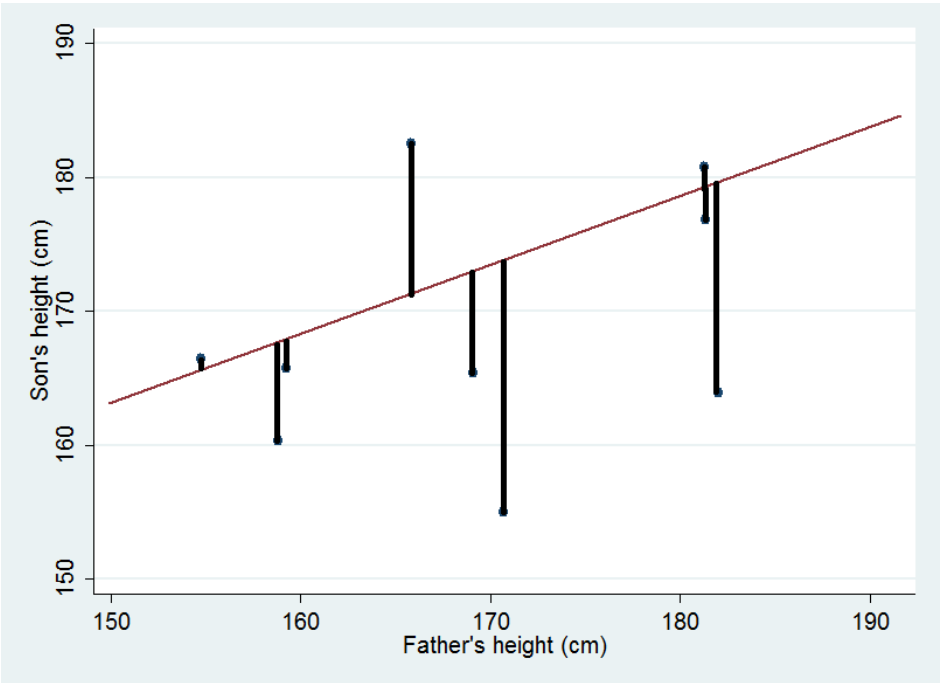
- Если $b > 0$, связь между x и y положительная
- Если $b < 0$, связь между x и y отрицательная
- Если $b = 0$, связь между x и y отсутствует
- b измеряет силу эффекта, но необходимо обращать внимание на шкалу, по которой измерена x

Предсказанные значения

- Для данных Галтона, $a = 86$ и $b = 0.51$, регрессионное уравнение принимает вид $y = 86 + 0.51x$
- Если мы знаем уравнение, мы с легкостью можем рассчитать предсказанное значение y на различных уровнях x
- Например, предсказанный рост мужчины, рост отца которого – 160 см: $\hat{y}_{(x=160)} = 86 + 0.51 * 160 = 167.6cm$
- Обозначение $\hat{y}$ используется для предсказанных значений

Регрессионные остатки (residuals, errors)

- Предсказанные значения y отличаются от реально наблюдаемых значений y
- Для наблюдения i, $e_i = y_i - \hat{y}_i$. e_i – остаток или ошибка
- Тогда для каждого наблюдения i, $y_i = a + bx_i + e_i$



Как провести регрессионную линию?

- Нам нужно знать a и b
- Нужно найти такие a и b, чтобы остатки были минимальными
- Мы не можем минимизировать $\sum e_i$ поскольку остатки могут быть положительными и отрицательными
- Мы можем минимизировать $\sum |e_i|$, но это математически менее удобно, чем
- ... минимизировать $\sum e_i^2$
- Отсюда – метод наименьших квадратов (МНК, OLS)

Регрессионные коэффициенты

- Решение:
  - $b = \frac{\sum(x_i - \bar{x})(y_i - \bar{y})}{\sum(x_i - \bar{x})^2}$
  - $a = \bar{y} - b\bar{x}$
  - где $\bar{y}$ – среднее значение y и $\bar{x}$ – среднее значение x
- Регрессионная линия всегда проходит через точку $(\bar{x}, \bar{y})$

Найдите регрессионные коэффициенты. Запишите регрессионное уравнение

x	y
1	5
2	2
6	8

Регрессия и корреляция

- r – коэффициент корреляции Пирсона
- $r = \frac{\sum(x_i - \bar{x})(y_i - \bar{y})}{\sqrt{\sum(x_i - \bar{x})^2}\sqrt{\sum(y_i - \bar{y})^2}}$
- Связь между коэффициентами регрессии и корреляции:
  - $b = r\frac{s_y}{s_x}$
  - s_y – стандартное отклонение y, s_x – стандартное отклонение x
- Если s_x=s_y, то $r = b$ и коэффициенты регрессии и корреляции равны друг другу
- Мы всегда можем стандартизировать переменную так, что средняя будет равна нулю, а стандартное отклонение – единице
  - $x_i^{st} = \frac{x_i - \bar{x}}{s_x}$

Регрессия и корреляция (2)

- И корреляция, и регрессия измеряют линейную связь. Если связь между двумя переменными не является линейной, эти коэффифиенты ее не отражают
- $r \in [0, 1]$
- Корреляция симметрична: $r(x, y) = r(y, x)$
- Регрессия не симметрична: $b_{x \to y} \neq b_{y \to x}$, если только не $s_x = s_y$
- Это НЕ означает, что в регрессии независимая переменная ВЛИЯЕТ на зависимую

r-квадрат

- В случае двух переменных квадрат коэффициента корреляции (r^2) называется коэффициентом детерминации
- Общая сумма квадратов (TSS) показывает, насколько далеко значения зависимой переменной y находятся от средней y: $TSS = \sum(y_i - \bar{y})^2$
- Сумма квадратов ошибок (SSE) показывает, насколько далеко значения зависимой переменной находятся от регрессионной линии: $SSE = \sum(y_i - \hat{y})^2$
- $r^2 = \frac{TSS - ESS}{TSS} = \frac{\sum(y_i - \bar{y})^2 - \sum(y_i - \hat{y})^2}{\sum(y_i - \bar{y})^2} = \frac{\text{“explained”variation}}{\text{totalvariation}}$
- r^2 интерпретируется как доля дисперсии зависимой переменной, “объясненной” независимой переменной

r-квадрат (2)

- $r^2 \in [0, 1]$. Если $r^2 = 0$, то связи между x и y нет. Если $r^2 = 1$, то все наблюдения находятся на регрессионной линии
- r^2 важен, если регрессионная модель используется с целью прогнозирования
- Однако важно понимать, что на самом деле регрессионные модели “объясняют” переменные исключительно в статистическом смысле слова

Регрессия в выборке и генеральной совокупности

- Пока мы говорили о том, как оценить регрессию в выборке
- О том, как распространить эту оценку на генеральную совокупность, мы будем говорить в следующий раз

Категориальные независимые переменные (факторы)

- Мы разобрали случай, в котором обе переменные являются интервальными
- Часто независимые переменные являются категориальными (пол, образование, место жительства)
- Если переменная дихотомическая, мы можем просто ее добавить в регрессионное уравение
- Если переменная не является дихотомической и имеет k значений, необходимо перекодировать ее в (k-1) дихотомические переменные со значениями 0 и 1 (фиктивные переменные, dummy variables)

Доход <- пол (СССР, 1991)

переменные	коэф.	ст.ошибка	p
мужчина (ref. женщина)	87	5	< 0.001
константа	169	4	< 0.001
n	2281		
R-квадрат	0.11		

Интерпретация коэффициентов

- Пол закодирован как 0 для женщин и 1 для мужчин
- Регрессионное уравнение: $\text{income} = 169 + 87 * \text{male}$
- Предсказание для мужчин: $\text{income} = 169 + 87 = 256$, для женщин – 169
- Регрессионный коэффициент (87) интерпретируется как разница в средних доходах между мужчинами и женщинами
- Статистически эта регрессия эквивалентна тесту на сравнению средних в двух группах (t-test)

Доход <- место жительства (СССР, 1991)

переменные	коэф.	ст.ошибка	p
Москва и Ленинград (ref.)			
Областные центры	-42	10	< 0.001
Другие города	-64	10	< 0.001
Село	-80	10	< 0.001
Константа	265	8	< 0.001
n	2281		
R-квадрат	0.03		

Интерпретация коэффициентов

- Регрессионное уравнение:
  $\text{income} = 265 - 42 * \text{reg.centre} - 64 * \text{oth.cities} - 80 * \text{country}$
- Коэффициенты показывают разницу средних доходах между Москвой/Ленинградом и областными центрами, Москвой/Ленинградом и другими городами, Москвой/Ленинградом и сельской местностью
- Статистически это фактически эквивалентно дисперсионному анализу (ANOVA)

Изменение базовой категории

переменные	коэф.	ст.ошибка	p
Другие города (ref.)			
Москва и Ленинград	64	9	< 0.001
Областные центры	22	7	< 0.01
Село	-16	7	< 0.05
Константа	201	5	< 0.001
n	2281		
R-квадрат	0.03		

Множественная регрессия

- До сих пор мы говорили о регрессии с одной независимой переменной
- Регрессия позволяет одновременно моделировать связь зависимой переменной с несколькими независимыми переменными
- На следующих лекциях мы будем говорить об этом подробнее

Доход <- возраст, пол, место жительства, образование (СССР, 1991)

переменные	коэф.	ст.ошибка	p
возраст	6.8	1.1	< 0.001
возраст в квадрате	-0.08	0.01	< 0.001
мужчины (ref.женщины)	79.4	4.9	< 0.001
место жительства(ref. Москва и Лениград)			
областные центры	-34.6	8.8	< 0.01
другие города	-50.2	8.5	< 0.001
село	-5.7	8.9	< 0.05
образование (ref. общее среднее)			
начальное	-9.3	12.1	0.44
незаконченное среднее	-7.5	7.4	0.31
среднее специальное	5.9	7.3	0.4
незаконченное высшее	-33.1	12.1	< 0.05
высшее	78.3	7.4	< 0.001
константа	82.3	22.8	< 0.001
n	2,274		
R-квадрат	0.24		

Интерпретация коэффициентов в множественной регрессии

- Коэффициент при каждой переменной интерпретируется так же как в обычной парной регрессии, но при условии, что все другие переменные “удерживаются” на одном уровне
- Коэффициент для высшего образования обозначает среднюю разницу в доходах между людьми с высшим и средним образованием, при этом мы сравниваем людей одного пола, возраста и живущих в похожих местах
- Подробнее на практических занятиях и на следующей лекции