Урок 10

Регрессия

10.1. Взаимосвязь нескольких признаков

В этом уроке будут рассмотрены методы, позволяющие проанализировать взаимосвязь между одним признаком и большим количеством других.

10.1.1. Пример исследования

Пусть исследователей интересует вопрос, влияет ли употребление алкоголя на успеваемость школьников. Лучший способ это проверить — провести эксперимент. Набирается случайная выборка школьников, и каждому из них назначается случайная еженедельная доза алкоголя. По окончании учебного года требуется измерить корреляцию между назначенной дозой и успеваемостью школьников.

Этот эксперимент идеален. Поскольку доза назначается случайно, выборка автоматически балансируется по всем возможным типам школьников, которые только могут быть. Этот эксперимент мог быть лучше только, если бы школьники сами не знали, какое количество алкоголя они принимают, но это достаточно сложно обеспечить.

Существенный недостаток этого эксперимента заключается в том, что его никогда не дадут провести — это неэтично. Такие ситуации возникают достаточно часто. Не получится исследовать взаимосвязь между уровнем насилия в видеоиграх и агрессивностью детей в жизни, поскольку нельзя заставить детей играть в видеоигры с высоким уровнем насилия какое-то продолжительное количество времени, если они сами этого не хотят. Иногда проведение эксперимента не только неэтично, а попросту невозможно. Например, если хочется понять, как влияет средняя дневная температура на вероятность возникновения лесного пожара, не существует никакого способа провести эксперимент, потому что средней дневной температурой в лесу управлять нельзя.

Единственное, что остается делать в условиях, когда нельзя провести эксперимент, — это использовать обзервационные данные, то есть данные, которые собраны путем наблюдения за выборкой. В задаче исследования успеваемости школьников можно, например, взять данные по 633 ученикам старших классов двух португальских школ, для которых известно большое количество разных демографических показателей, в том числе, успеваемость. В частности, среди всех показателей есть уровень потребления алкоголя по выходным и финальная оценка по португальскому языку.

На рисунке 10.1 изображены эти два показателя для 633 учащихся. Видно, что эти две величины друг с другом отрицательно скоррелированы. Эта корреляция значима. Возникает вопрос: значит ли это, что потребление алкоголя влияет на успеваемость старшеклассников, или что чем больше алкоголя они потребляют, тем хуже они учатся. Чтобы точнее ответить на него, можно использовать еще 29 признаков, которые есть в наборе данных о школьниках. Эти признаки потенциально влияют на успеваемость гораздо сильнее, чем употребление алкоголя. Например, возраст учеников или доход их родителей могут определять успеваемость гораздо более явно. Теперь требуется узнать, останется ли у потребления алкоголя предсказательная сила при учёте остальных признаков, и можно ли утверждать, что потребление алкоголя вызывает снижение оценки по португальскому языку, то есть ли причинно-следственная связь между этими двумя признаками.

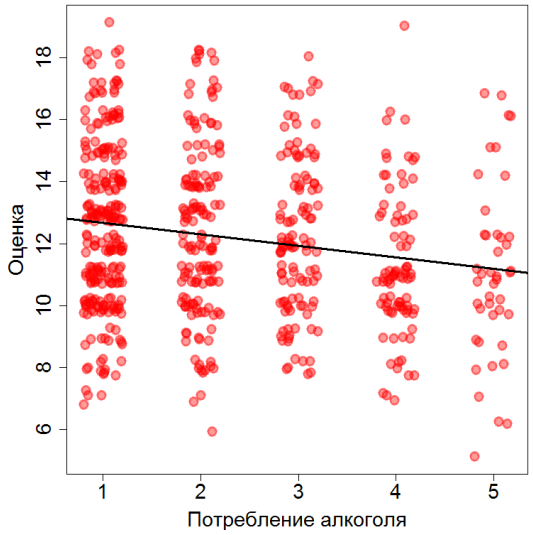

Рис. 10.1: Данные об уровне употребления алкоголя по выходным и финальной оценке по португальскому языку у учащихся двух португальских школ

10.1.2. Линейная регрессия

Оказывается, на такие вопросы можно отвечать с помощью линейной регрессии. Задача линейной регрессии: есть n объектов, на которых измерены значения k признаков $x_1, \ldots, x_k$, и, кроме того, для них известно значение отклика y. Требуется найти вектор констант β такой, что:

$$y \approx \beta x.$$

При построении регрессии строится наилучшее линейное по x приближение условного математического ожидания y при таких x:

$$\mathbb{E}\left(y \,|x\right) \approx \beta_0 + \sum_{j=1}^{k} \beta_j x_j.$$

В линейной регрессии коэффициент β_j показывает, насколько в среднем увеличивается отклик y, если x_j увеличивается на 1, а все остальные x зафиксированы. Таким образом, используя регрессию, можно изолировать эффект интересующей переменной и посмотреть на него отдельно. Иногда этот эффект можно даже интерпретировать как причинно-следственную связь, при выполнении некоторых специальных условий. Строить обычную линейную регрессию очень просто. Однако если по построенной модели хочется делать какие-то выводы с использованием статистических методов, необходимо приложить дополнительные усилия. Именно этому и будет посвящен урок.

10.2. Свойства решения задачи

10.2.1. Задача линейной регрессии

Итак, решается задача линейной регрессии:

$$\mathbb{E}\left(y \,|x\right) \approx \beta_0 + \sum_{j=1}^{k} \beta_j x_j.$$

Для того, чтобы больше не думать про коэффициент b_0, можно добавить в матрицу объекты-признаки X единичный столбец:

$$X = \begin{pmatrix} x_{10} = 1 & x_{11} & \ldots & x_{1k} \\ \vdots & \vdots & \ddots & \vdots \\ x_{n0} = 1 & x_{n1} & \ldots & x_{nk} \end{pmatrix} \quad y = \begin{pmatrix} y_1 \\ \vdots \\ y_n \end{pmatrix} \quad \beta = \begin{pmatrix} \beta_0 \\ \beta_1 \\ \vdots \\ \beta_k \end{pmatrix}$$

Теперь эта матрица размера $n \times (k+1)$.

Задача регрессии будет решаться методом наименьших квадратов без использования регуляризаторов:

$$\|y - X\beta\|_2^2 \to \min_\beta .$$

Точное решение этой задачи известно, $\hat{\beta}$ выражается аналитически:

$$\hat{\beta} = \left(X^T X\right)^{-1} X^T y.$$

Можно посчитать и $\hat{y}$, то есть предсказание модели на объектах, на которых она обучается:

$$\hat{y} = X\left(X^T X\right)^{-1} X^T y.$$

Чтобы найти качество решения, полученного методом наименьших квадратов, определим величину TSS (Total Sum of Squares) — разброс y относительно своего среднего:

$$\mathrm{TSS} = \sum_{i=1}^{n} (y_i - \bar{y})^2 .$$

Оказывается, что этот разброс можно поделить на две части:

$$\mathrm{TSS} = \mathrm{ESS} + \mathrm{RSS} .$$

Одна из частей, объясненная сумма квадратов, — это сумма квадратов отклонений среднего y от предсказанных y:

$$\mathrm{ESS} = \sum_{i=1}^{n} (\hat{y}_i - \bar{y})^2 .$$

Вторая часть, остаточная сумма квадратов, — это сумма квадратов отклонений предсказанных y от их истинных значений:

$$\mathrm{RSS} = \sum_{i=1}^{n} (y_i - \hat{y}_i)^2 .$$

По этим величинам, ESS и TSS, можно составить меру R^2, которая называется коэффициентом детерминации:

$$R^2 = \frac{\mathrm{ESS}}{\mathrm{TSS}} = 1 - \frac{\mathrm{RSS}}{\mathrm{TSS}}.$$

По сути, это доля объясненной дисперсии отклика во всей дисперсии отклика.

10.2.2. Предположения МНК

Для того, чтобы решение метода наименьших квадратов обладало интересующими нас свойствами, необходимо сделать следующие предположения.

- Предполагается, что истинная модель y действительно линейна:

  $$y = X\beta + \varepsilon,$$

  где ε — это какая-то ошибка.

- Предполагается, что наблюдения, по которым оценивается модель, случайны, то есть объекты дают независимую выборку наблюдений (x_i, y_i).

- Предполагается, что матрица X — матрица полного столбцового ранга:

  $$\operatorname{rank} X = k + 1,$$

  то есть ни один из признаков не должен являться линейной комбинацией других. Поскольку среди столбцов есть константа, никакой из признаков в выборке не должен быть константой.

- Предполагается, что ошибка случайна:

$$\mathbb{E}(\varepsilon|x) = 0.$$

Уже из этих четырех предположений можно вывести полезное свойство оценок, получаемых методом наименьших квадратов. Если они выполняются, то оценки $\hat{\beta}$ являются несмещенными

$$\mathbb{E}\hat{\beta}_j = \beta_j$$

и состоятельными оценками истинных β:

$$\forall \gamma > 0 \lim_{n\to\infty} P\left(\left|\beta_j - \hat{\beta}_j\right| < \gamma\right) = 1.$$

К четырем предположениям можно добавить еще пятое — предположение гомоскедастичности ошибок.

- Предполагается, что дисперсия ошибки не зависит от значений признака:

$$\mathbb{D}(\varepsilon|x) = \sigma^2.$$

Вместе эти пять предположений называются предположениями Гаусса-Маркова. Теорема Гаусса-Маркова утверждает, что если эти предположения выполняются, то МНК-оценки имеют наименьшую дисперсию в классе всех оценок β, линейных по y. То есть оценки методом наименьших квадратов при выполнении этих пяти предположений в каком-то смысле являются оптимальными.

Из сделанных предположений вытекает следующее выражение для дисперсии МНК-оценок:

$$\mathbb{D}\left(\hat{\beta}_j\right) = \frac{\sigma^2}{TSS_j\left(1 - R_j^2\right)},$$

то есть:

- чем больше σ^2, тем больше дисперсия $\hat{\beta}_j$;
- чем больше вариация значений x_j в выборке, тем меньше дисперсия $\hat{\beta}_j$;
- чем лучше признак x_j объясняется линейной комбинацией оставшихся признаков, тем больше дисперсия $\hat{\beta}_j$.

По предположению о полноте столбцового ранга матрицы X коэффициент детерминации $R_j^2 < 1$, но, тем не менее, может быть $R_j^2 \approx 1$. Такая ситуация называется мультиколлинеарностью.

В матричном виде выражение для дисперсии вектора $\hat{\beta}$ выглядит вот так:

$$\mathbb{D}\left(\hat{\beta}\right) = \sigma^2\left(X^T X\right)^{-1}.$$

Если матрица X содержит столбцы, которые почти линейно зависимы, то матрица $X^T X$ будет плохо обусловлена. При обращении этой матрицы будет получаться численная неустойчивость, поэтому дисперсия оценок $\hat{\beta}_j$ будет велика.

Следует обратить внимание, что определение «мультиколлинеарности» не включает случай, когда столбцы полностью линейно зависимы. Мультиколлинеарность — это близкая к линейной зависимость признаков.

К 5 предположениям Гаусса-Маркова можно добавить еще одно, предположение о нормальности ошибки ε:

$$\varepsilon|x \sim N\left(0, \sigma^2\right).$$

Это эквивалентно следующей записи:

$$y|x \sim N\left(x\beta, \sigma^2\right).$$

Если выполняются эти 6 предположений, то оценки, даваемые методом наименьших квадратов, совпадают с оценками максимального правдоподобия. Это открывает доступ к прекрасным свойствам оценок максимального правдоподобия. Из этих 6 предположений вытекает, что оценки метода наименьших квадратов, во-первых, имеют наименьшую дисперсию среди всех несмещенных оценок β. Во-вторых, имеют нормальное распределение:

$$N\left(\beta, \sigma^2\left(X^T X\right)^{-1}\right).$$

Далее, дисперсию шума σ^2 можно оценить с помощью RSS:

$$\hat{\sigma}^2 = \frac{RSS}{n-k-1}.$$

Кроме того, отношение RSS к истинной дисперсии будет распределено по χ^2:

$$\frac{\text{RSS}}{\sigma^2} \sim \chi^2_{n-k-1}.$$

Наконец, следующее очень сильное свойство. Для любого вещественного вектора c длины $k+1$ справедливо следующее утверждение:

$$\frac{c^T\left(\beta - \hat{\beta}\right)}{\hat{\sigma}\sqrt{c^T\left(X^TX\right)^{-1}c}} \sim St(n-k-1).$$

10.2.3. Последствия

Если выполняются описанные предположения, то можно строить доверительный интервалы для коэффициентов β_j, доверительные интервалы для среднего отклика $\mathbb{E}\left(y\,|x\right)$ и предсказательные интервалы для значения $y\,|x$ Далее будет описано, как всё это делать.

10.3. Интервалы и гипотезы

10.3.1. Построение доверительных и предсказательных интервалов

В предыдущей части утверждалось, что, если выполняются шесть необходимых предположений, из этого вытекают очень полезные свойства. Эти свойства можно немедленно использовать. Во-первых, $100(1-\alpha)\%$ доверительный интервал для дисперсии шума σ^2 можно построить через отношение RSS к квантилям распределения χ^2:

$$\frac{\text{RSS}}{\chi^2_{n-k-1,1-\alpha/2}} \leq \sigma^2 \leq \frac{\text{RSS}}{\chi^2_{n-k-1,\alpha/2}}.$$

Во-вторых, чтобы построить доверительные интервалы для коэффициента β_j, можно использовать последнее утверждение (о распределении Стьюдента), и в качестве вектора c выбрать вектор, состоящий из всех нулей, в котором на j позиции стоит 1 $c = \left(0\ldots\underset{j}{1}0\ldots0\right)$. Тогда $100(1-\alpha)\%$ доверительный интервал для коэффициента β_j задаётся следующим образом:

$$\hat{\beta}_j \pm t_{n-k-1,1-\alpha/2}\hat{\sigma}\sqrt{(X^TX)^{-1}_{jj}}.$$

Чтобы построить доверительный интервал для математического ожидания отклика y на новом объекте, задаваемом вектором x_0, в качестве вектора c можно использовать x_0. $100(1-\alpha)\%$ доверительный интервал для $\mathbb{E}\left(y\,|x_0\right)$ готов:

$$x_0^T\hat{\beta} \pm t_{n-k-1,1-\alpha/2}\hat{\sigma}\sqrt{x_0^T\left(X^TX\right)^{-1}x_0}.$$

Чтобы построить предсказательный интервал для значения отклика на этом же самом объекте $y\left(x_0\right) = x_0^T\beta + \varepsilon\left(x_0\right)$, необходим дополнительно учесть ещё дисперсию ошибки:

$$x_0^T\hat{\beta} \pm t_{n-k-1,1-\alpha/2}\hat{\sigma}\sqrt{1 + x_0^T\left(X^TX\right)^{-1}x_0}.$$

Формула для предсказательного интервала отличается от формулы для доверительного интервала условного математического ожидания только единицей, стоящей под корнем.

10.3.2. Критерий Стьюдента

Для проверки гипотезы

$$H_0 \colon \beta_j = 0$$

можно использовать T-критерий Стьюдента (таблица 10.1). Гипотеза о равенстве нулю коэффициента β_j означает, что признак x_j не влияет на отклик y.

нулевая гипотеза:	$H_0 \colon \beta_j = 0;$
альтернатива:	$H_1 \colon \beta_j <\neq> 0;$
статистика:	$T = \frac{\hat{\beta}_j}{\sqrt{\frac{\text{RSS}}{n-k-1}(X^T X)^{-1}_{jj}}};$
нулевое распределение:	$T \sim St(n-k-1).$

Таблица 10.1: Описание t-критерия Стьюдента

Если справедлива нулевая гипотеза, статистика данного критерия имеет распределение Стьюдента с числом степеней свободы $n-k-1$ (рисунок 10.2).

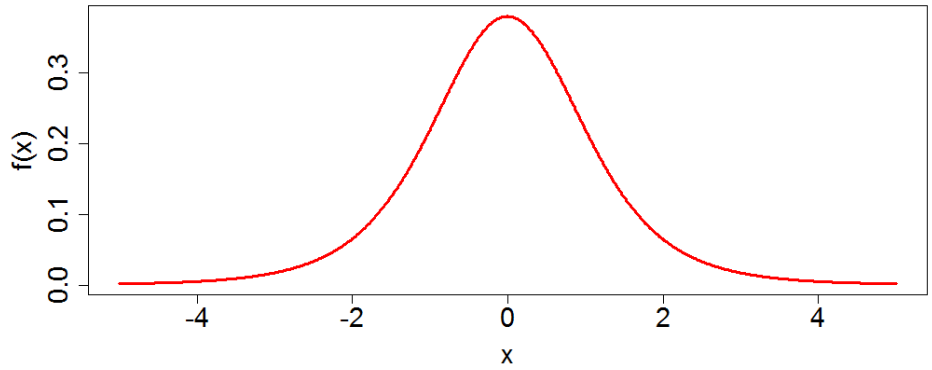

Рис. 10.2: Распределение Стьюдента

10.3.3. Пример

Пусть есть 12 испытуемых, и x — это результат прохождения ими составного теста на скорость реакции, а y — это их результат на симуляторе транспортного средства. Значение y получать долго и дорого, поэтому поставлена задача предсказания y по x. Необходимо понять, можно ли это делать.

Строится регрессионная модель

$$y = \beta_0 + \beta_1 x + \varepsilon.$$

Затем проверяется, что переменная x значима для предсказания y. Нулевая гипотеза

$$H_0 \colon \beta_1 = 0$$

против двусторонней альтернативы

$$H_1 \colon \beta_1 \neq 0$$

критерием Стьюдента отвергается. Достигаемый уровень значимости $p = 2.2021 \times 10^{-5}$.

10.3.4. Критерий Фишера

Для проверки гипотезы о том, что сразу несколько коэффициентов модели равны 0, будет использоваться не критерий Стьюдента, а критерий Фишера (таблица 10.2).

Матрицу объекты-признаки X нужно поделить на две части:

$$\underset{n\times(k+1)}{X} = \left(\underset{n\times(k+1-k_1)}{X_1}, \underset{n\times k_1}{X_2} \right).$$

нулевая гипотеза:	$H_0 \colon \beta_2 = 0;$
альтернатива:	$H_1 \colon H_0$ неверна;
статистика:	$\mathrm{RSS}_r = \|y - X_1\beta_1\|_2^2,$
	$\mathrm{RSS}_{ur} = \|y - X\beta\|_2^2,$
	$F = \frac{(\mathrm{RSS}_r - \mathrm{RSS}_{ur})/k_1}{\mathrm{RSS}_{ur}/(n-k-1)};$
нулевое распределение:	$F \sim F(k_1, n-k-1).$

Таблица 10.2: Описание критерия Фишера

В первую часть X_1 помещаются все признаки, которые мы хотим оставить в модели (константу нужно обязательно оставить там же). Во вторую часть X_2 переносят все признаки, для которых требуется проверить гипотезу о значимости влияния на отклик. За β_1 и β_2 обозначаются соответствующие куски вектора параметров модели β:

$$\underset{(k+1)\times 1}{\beta^T} = \left(\underset{(k+1-k_1)\times 1}{\beta_1^T}, \underset{k_1\times 1}{\beta_2^T} \right)^T.$$

Проверяется нулевая гипотеза о том, что все компоненты вектора β_2 равны нулю. Это делается с помощью статистики F, которая определяется через соотношение двух RSS, где RSS_r — это RSS сокращённой модели (модель, в которой признаки из X_2 вообще не используются), а RSS_{ur} — это RSS полной модели, в которой есть признаки и X_1, и X_2.

Если нулевая гипотеза справедлива, то такая статистика F, составленная из двух RSS, имеет распределение Фишера с числом степеней свободы k_1 и $n-k-1$ (рисунок 10.3).

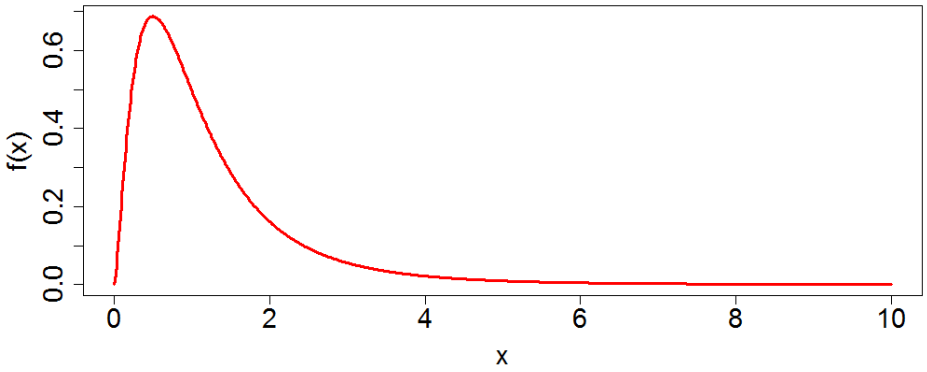

Рис. 10.3: Распределение Фишера

10.3.5. Пример

Пусть есть 1191 детей, для которых известны: их вес при рождении *weight*, среднее число сигарет, которые выкуривала мать за один день беременности *cigs*, номер ребёнка у матери *parity*, среднемесячный доход семьи *inc*, а также длительность получения образования в годах матерью *med* и отцом *fed*. По этим данным строится модель:

$$weight = \beta_0 + \beta_1 cigs + \beta_2 parity + \beta_3 inc + \beta_4 med + \beta_5 fed + \varepsilon.$$

Требуется проверить гипотезу о том, что образование родителей не является значимым предиктором при предсказании веса ребёнка при рождении. Для этого используется критерий Фишера. Проверяется нулевая гипотеза

$$H_0 \colon \beta_4 = \beta_5 = 0$$

против общей альтернативы:

$$H_1 \colon H_0 \text{ неверна.}$$

Критерий Фишера даёт достигаемый уровень значимости $p = 0.2421$, то есть данные не позволяют отклонить нулевую гипотезу.

10.3.6. Критерии Фишера и Стьюдента

Если $k_1 = 1$, то критерий Фишера даёт абсолютно такой же достигаемый уровень значимости, какой дал бы критерий Стьюдента для этого же самого признака при использовании двусторонней альтернативы.

Если $k_1 > 1$, могут возникать разные неоднозначные ситуации. Например, критерий Фишера может говорить, что гипотеза незначимости признаков X_2 отвергается. При этом критерий Стьюдента не может отвергнуть никакую из гипотез о признаках, лежащих внутри X_2. Получается странная ситуация: все вместе признаки значимо определяют отклик, но отдельно ни один из них значимо на отклик не влияет. Это можно объяснить двумя способами. Во-первых, такая ситуация может возникать, если отдельные признаки из X_2 недостаточно хорошо объясняют отклик, но их совокупный эффект при прогнозировании y значим. Во-вторых, признаки из X_2 могут быть мультиколлинеарны. Мультиколлинеарность приводит к численной неустойчивости критериев Стьюдента и Фишера, поэтому их достигаемые уровни значимости могут быть неадекватными.

Теперь противоположная ситуация. Пусть критерий Фишера не отвергает гипотезу о незначимости признаков из X_2, а критерий Стьюдента по отдельным компонентам X_2 какие-то из гипотез отвергает. То есть все вместе признаки незначимы, а какие-то из них по отдельности оказываются значимыми. Для этого тоже может быть два объяснения. Первый вариант: незначимые признаки из X_2 маскируют влияние значимых. Второй вариант: значимость отдельных признаков из X_2 — это результат эффекта множественной проверки гипотез. Действительно, критерии Фишера проверяют всего одну гипотезу, а критерии Стьюдента проверяют целую серию из k_1 гипотез, и какие-то из них могут отклониться просто случайно.

10.3.7. Критерий Фишера для проверки гипотезы о незначимости всех признаков

Критерий Фишера имеет особенный вид (таблица 10.3), если требуется проверить гипотезу о том, что все признаки X для предсказания y не нужны, то есть лучшее предсказание для y — это константа.

нулевая гипотеза:	$H_0\colon \beta_1 = \cdots = \beta_k = 0;$
альтернатива:	$H_1\colon H_0$ неверна;
статистика:	$F = \frac{R^2/k}{(1-R^2)/(n-k-1)};$
нулевое распределение:	$F \sim F(k, n-k-1).$

Таблица 10.3: Описание критерия Фишера для проверки гипотезы о незначимости всех признаков

Нулевое распределение статистики точно такое же, как и раньше, — это распределение Фишера (рисунок 10.3).

Пример. В предыдущей задаче о весе детей при рождении можно проверить гипотезу о том, что построенная модель вообще имеет хоть какой-то смысл. Проверяем гипотезу о том, что все β равны нулю:

$$H_0\colon \beta_1 = \cdots = \beta_5 = 0$$

против общей альтернативы

$$H_1\colon H_0 \text{ неверна.}$$

Критерием Фишера нулевая гипотеза уверено отвергается, достигаемый уровень значимости $p = 6 \times 10^{-9}$.

10.4. Проверка предположений

В этой части пойдёт речь о том, как проверять шесть предположений, лежащих в основе всей статистической машины, с помощью которой проверяется значимость коэффициентов регрессии.

10.4.1. Линейность отклика

Первое предположение — это предположение о линейности отклика. Утверждается, что y в действительности представляет собой линейную комбинацию X с какой-то случайной ошибкой ε:

$$y = X\beta + \varepsilon.$$

Естественно, это предположение в точности не выполняется никогда. Трудно ожидать, что отклик y в действительности — это линейная комбинация рассматриваемых признаков x. Линейная модель, как и все остальные, неверна, но очень полезна, и кроме того, устойчива к небольшим отклонениям от линейности. Поэтому единственное, что требуется проверить, — это нет ли каких-то огромных отклонений от линейности y по x. Чтобы убедиться в отсутствии больших отклонений от линейнойсти, нужно анализировать остатки:

$$\hat{\varepsilon}_i = y_i - \hat{y}_i,\ i = 1, \ldots, n,$$

где y — это истинные значения, а $\hat{y}$ — предсказываемые.

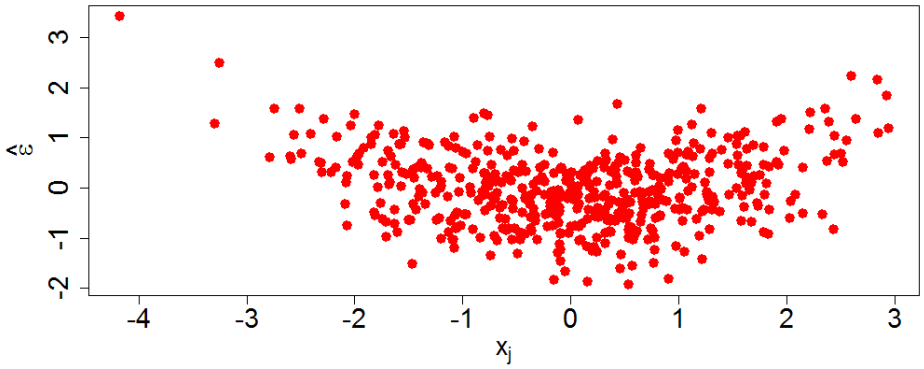

Рис. 10.4: График для проверки отклонений от линейности. По горизонтальной оси — значения признаков, по вертикальной — остатки.

Для остатков необходимо построить графики. По горизонтальной оси откладывается значение каждого из признаков x_j, по вертикальной оси — остатки, и нужно смотреть, как выглядит получающееся облако точек. Если, например, оно выглядит как на рисунке 10.4, представляет собой какую-то параболу, то, скорее всего это значит, что отклик y зависит от квадрата признака x_j. Такую зависимость можно учесть, просто добавив в матрицу X столбец, соответствующий x_j^2.

На таком графике можно обнаруживать и другие осмысленные функциональные зависимости. Если такие зависимости видны, нужно просто добавить в матрицу X соответствующий столбец.

10.4.2. Случайность выборки

Следующее предположение — это предположение о случайности выборки. Требуется, чтобы выборка была независимой и одинаково распределенной. Это предположение может нарушаться несколькими способами. Первый способ более тяжелый: если объекты, на которых измерены признаки и отклик, зависимы, то всё плохо: дисперсии ошибки и коэффициентов недооцениваются, и все статистические критерии, которые на этом основаны, перестают работать корректно.

Еще это предположение может нарушаться, если выборка отобрана из генеральной совокупности не случайно, а каким-то образом отфильтрована. Фильтровать генеральную совокупность по какому-то признаку z можно только в случае, если

$$\mathbb{E}\,(y\,|x, z\,) = \mathbb{E}\,(y\,|x\,)\,,$$

то есть z не добавляет никакой новой информации об y.

Если выборка отфильтровывалась как-то иначе, например, просто по одному из признаков, содержащихся в x, то выводы, построенные по такой модели, можно распространять только на отфильтрованную генеральную совокупность. Например, если в выборке испытуемые только младше 50 лет, то нельзя ничего сказать об испытуемых в генеральной совокупности, которым больше 50 лет.

10.4.3. Полнота ранга X

Следующее предположение: матрица X должна иметь полный столбцовый ранг, то есть

$$\operatorname{rank} X = k + 1.$$

Если в выборке есть линейно зависимые признаки, то дисперсия оценки коэффициентов при таких признаках будет бесконечной. Это не очень удобно при построении доверительных интервалов: они будут иметь бесконечную ширину. И кроме того, гипотезы тоже так не проверить.

Если возникла такая проблема, это значит, что от каких-то признаков в модели придется избавиться. Помимо всего прочего, для категориальных переменных нельзя использовать one-hot encoding, которое использовалось в предыдущих курсах. Дело в том, что при кодировании каждого уровня фактора своей бинарной переменной мы получаем, что в сумме такие переменные дают единичный столбец, а он в матрице X уже есть, поэтому столбцы получаются линейно зависимы. Вместо этого нужно использовать другой способ кодирования: dummy-кодирование. Если признак x_j принимает m различных значений, то его нужно кодировать $m-1$ фиктивной переменной.

Тип должности	x_1	x_2
рабочий	0	0
инженер	1	0
управляющий	0	1

Таблица 10.4: Dummy-кодирование

Пусть y — это уровень заработной платы, а x — это занимаемая человеком должность: рабочий, инженер или управляющий. Эти три значения будут кодироваться двумя фиктивными переменными: x_1 и x_2 (таблица 10.4). В полученной регрессионной модели два признака: x_1 и x_2:

$$y = \beta_0 + \beta_1 x_1 + \beta_2 x_2.$$

Коэффициенты β_1 и β_2 при них кодируют среднюю разницу в уровнях зарплат инженера и рабочего и управляющего и рабочего. В регрессионных моделях с использованием dummy-кодированиея интерпретация коэффициентов β модели всегда ведется относительно уровня фактора, который закодирован всеми 0. Можно менять кодировку dummy, используя соображения о том, какая из моделей будет удобнее интерпретироваться.

10.4.4. Случайность ошибок

На очереди предположение о случайности ошибки:

$$\mathbb{E}(\varepsilon | x) = 0.$$

Гипотезу

$$H_0 \colon \mathbb{E}(\varepsilon | x) = 0$$

можно очень легко проверить по данным. Для этого нужно построить регрессию y по x, вычислить остатки и проверить гипотезу о том, что среднее значение остатков равно 0. Это можно сделать, например, с помощью критерия Стьюдента.

10.4.5. Гомоскедастичность ошибок

Пятое предположение — предположение гомоскедастичности ошибки:

$$\mathbb{D}(\varepsilon | x) = \sigma^2.$$

Это предположение можно проверять двумя способами. Первый, нестрогий, — это визуальный анализ. Нужно построить графики зависимости остатков от всех признаков x_j (рисунок 10.5) и посмотреть, выглядят ли точки на этом графике как горизонтальная полоса. Если вместо горизонтальной полосы на графике изображено что-то расширяющееся или сужающееся, значит, предположение гомоскедастичности не выполняется.

Формально это предположение можно проверять с помощью критерия Бройша-Пагана (таблица 10.5).

нулевая гипотеза:	$H_0 \colon \mathbb{D}\varepsilon = \sigma^2$;
альтернатива:	$H_1 \colon H_0$ неверна;
статистика:	$LM = nR^2_{\hat{\varepsilon}^2}$, $R^2_{\hat{\varepsilon}^2}$ — коэффициент детерминации при регрессии $\hat{\varepsilon}^2$ на x;
нулевое распределение:	$LM \sim \chi^2_k$.

Таблица 10.5: Описание критерия Бройша-Пагана

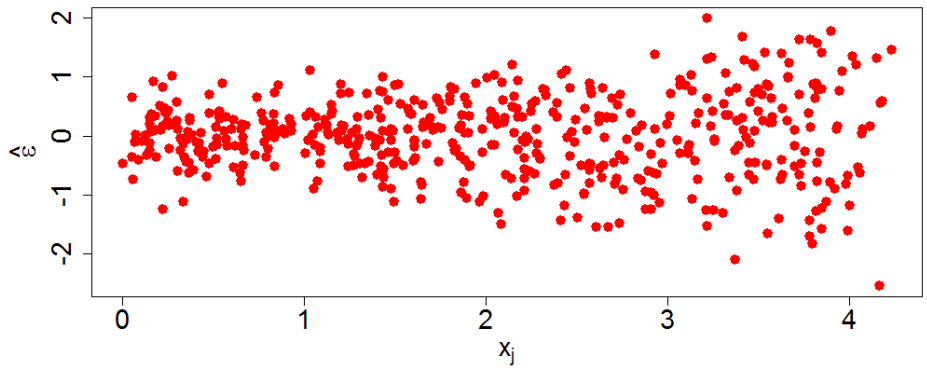

Рис. 10.5: График зависимости остатков от значения признаков x_j

Если справедлива нулевая гипотеза, статистика этого критерия имеет распределение хи-квадрат с числом степеней свободы k (рисунок 10.6).

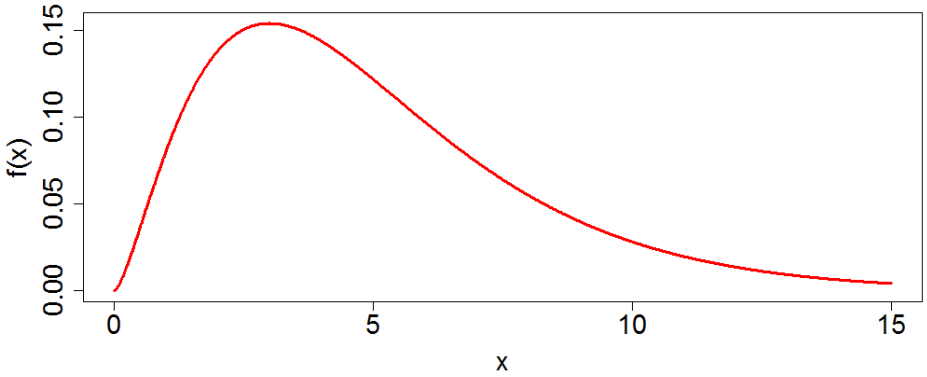

Рис. 10.6: Распределение хи-квадрат

10.4.6. Нормальность ошибок

Наконец, шестое предположение — это предположение нормальности. Способы проверки нормальности уже разбирались ранее. Есть визуальный способ: нужно построить ку-ку график и посмотреть, лежат ли точки на этом графике более-менее на одной прямой. Также есть формальный способ: можно использовать статистические критерии для проверки нормальности. Среди всего разнообразия критериев рекомендуется использовать критерий Шапиро-Уилка.

10.5. Регрессия и причинно-следственные связи

10.5.1. Упражнения и холестерин

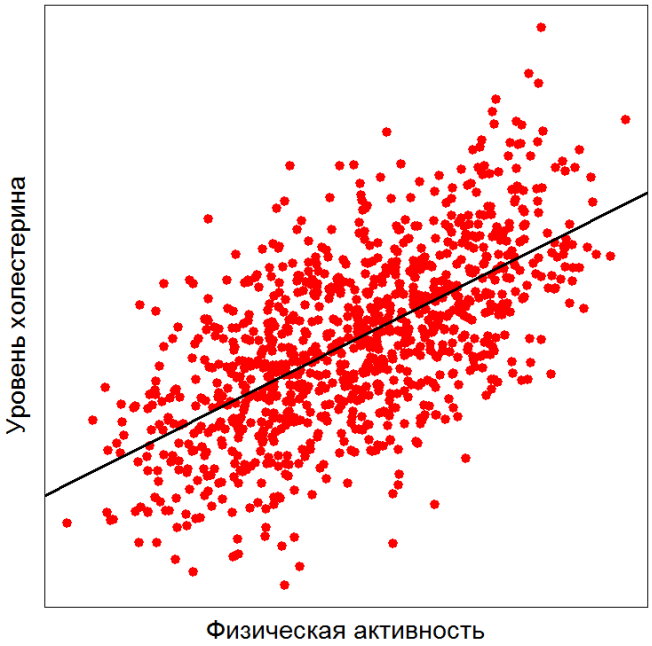

Рис. 10.7: Данные об уровне физической активности (по горизонтальной оси) и уровне холестерина в крови (по вертикальной оси)

Проводится исследование о связи уровня физической активности человека и уровня холестерина у него в крови, участвуют 10000 испытуемых. На рисунке 10.7 на графике по горизонтальной оси отложен уровень физической активности, по вертикальной — уровень холестерина. Видно, что эти два признака положительно коррелированы, поскольку облако вытянуто вдоль диагонали.

Можно проверить гипотезу о том, что по уровню физической активности ex можно каким-то образом предсказывать уровень холестерина $chol$. Для этого нужно построить регрессионную модель с одним-единственным признаком

$$chol = \beta_0 + \beta_1 ex$$

и проверить гипотезу

$$H_0 \colon \beta_1 = 0$$

против альтернативы

$$H_0 \colon \beta_1 > 0.$$

Критерий Стьюдента говорит, что нулевая гипотеза отвергается против этой альтернативы с очень маленьким достигаемым уровнем значимости $p = 2 \times 10^{-16}$. Даже если мы бы альтернатива была двухсторонней, получился бы достигаемый уровень значимости был бы $p = 2 \times 10^{-16}$ — это тоже мало.

Стоит посмотреть, как эти же самые данные выглядят в разрезе возраста испытуемых. На рисунке 10.8 размечены пять возрастных групп: от левого нижнего угла к верхнему правому располагаются группы 10–20, 20–30, 30–40, 40–50 и 50–60 лет. В каждой возрастной группе уровень холестерина и количество физических упражнений друг с другом связаны отрицательно, но при этом в каждой следующей группе оба признака — и уровень холестерина, и уровень физической активности — растут с возрастом.

Теперь можно построить линейную регрессионную модель с двумя признаками: уровень холестерина $chol$ будет предсказываться по возрасту age и количеству физических упражнений ex:

$$chol = \beta_0 + \beta_1 ex + \beta_2 age$$

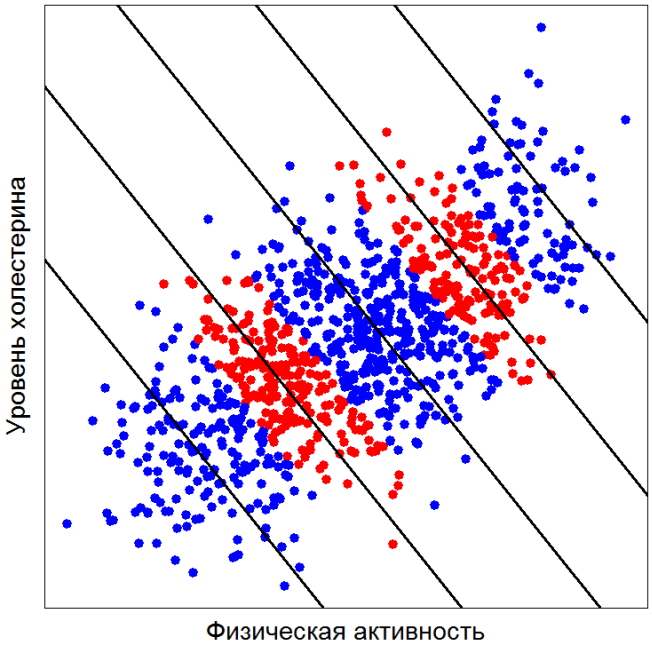

Рис. 10.8: Данные об уровне физической активности (по горизонтальной оси) и уровне холестерина в крови (по вертикальной оси) для разных возрастных групп

Проверяется утверждение, что количество физических упражнений хорошо предсказывает уровень холестерина. Гипотеза

$$H_0 \colon \beta_1 = 0$$

будет проверяться против альтернативы

$$H_0 \colon \beta_1 < 0.$$

Критерий Стьюдента дает достигаемый уровень значимости такой же маленький, как и раньше: $p = 2 \times 10^{-16}$, то есть нулевая гипотеза снова отвергается.

Итак, есть две модели. Первая модель показывает, что физические упражнения положительно влияют на уровень холестерина (то есть, чем больше вы упражняетесь, тем больше холестерина у вас в крови). Вторая модель демонстрирует ровно противоположное: если известен возраст человека, то чем больше он упражняется, тем ниже уровень холестерина у него в крови. Нужно решить, какой из этих двух выводов принять как финальный. В этой задаче можно включить здравый смысл и понять, что именно вторая модель верна. Кажется, что физические упражнения улучшают здоровье, поэтому, наверное, уровень холестерина должен снижаться. Но не во всех задачах такая опция доступна, стоит попробовать не использовать здравый смысл. Можно рассуждать так: вторая модель более подробна, она содержит больше признаков, значит, она богаче и, возможно, благодаря этому ее выводы более правильные. То есть модель, в которой больше признаков, лучше, чем модель, в которой их меньше.

10.5.2. Средний балл и мотивация

Пусть теперь первый признак — это средний балл выпускника из школы, а второй — это результат выпускника на мотивационном тесте во время собеседования при поступлении в вуз. Если посмотреть на облако точек, которое показано на рисунке 10.9, то кажется, что эти два признака вообще никак не связаны друг с другом. Красные точки на графике — это школьники, которые поступили в вуз, а синие — это те, которые не поступили. Видно, что правила приема в вуз устроены достаточно просто: поступают ученики, у которых или высокий средний балл, или хорошие результаты на тесте по мотивации. Требуется понять, влияет ли на результаты теста по мотивации средний балл. По конфигурации облака точек кажется, что не влияет, но это можно проверить формально: построить простую регрессионную модель, предсказывающую результат теста на мотивацию mot по среднему баллу SAT:

$$mot = \beta_0 + \beta_1 SAT.$$

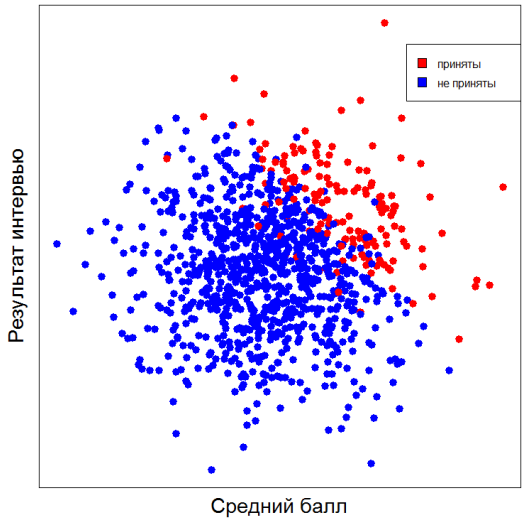

Рис. 10.9: Данные о среднем балле (по горизонтальной оси) и мотивации (по вертикальной оси). Красные точки — школьники, поступившие в вуз, синие — не поступившие.

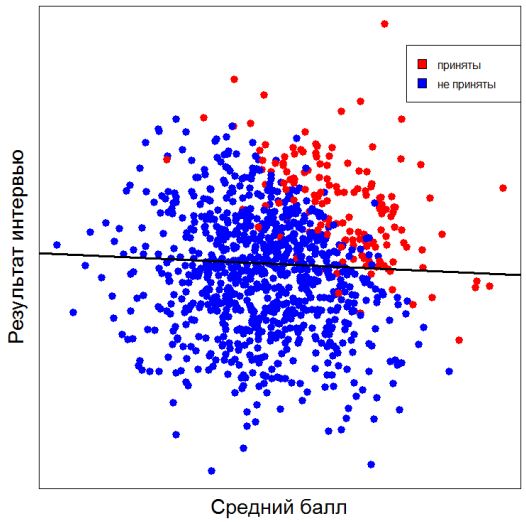

Рис. 10.10: Данные о среднем балле (по горизонтальной оси) и мотивации (по вертикальной оси) и прямая,соответствующая регрессии, которая предсказывает результат теста на мотивацию по среднему баллу

Далее нужно проверить гипотезу

$$H_0 \colon \beta_1 = 0$$

против двухсторонней альтернативы

$$H_0 \colon \beta_1 \neq 0$$

Критерий Стьюдента дает достигаемый уровень значимости $p = 0.1452$, нулевую гипотезу отвергнуть не получается, то есть нельзя утверждать, что средний балл влияет на результат теста по мотивации.

В эту регрессионную модель можно добавить еще один признак *acc*: «поступил ли человек в вуз» (10.11):

$$mot = \beta_0 + \beta_1 SAT + \beta_2 acc$$

В такой регрессионной модели снова проверяется нулевая гипотеза

$$H_0 \colon \beta_1 = 0$$

против двухсторонней альтернативы

$$H_0 \colon \beta_1 \neq 0.$$

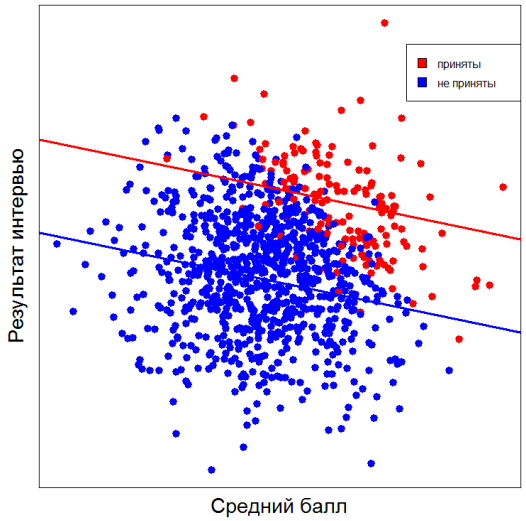

Рис. 10.11: Данные о среднем балле (по горизонтальной оси) и мотивации (по вертикальной оси) и прямая,соответствующая регрессии, которая предсказывает результат теста на мотивацию по среднему баллу и по тому, поступил ли человек в вуз

На этот раз критерий Стьюдента уверенно отвергает эту нулевую гипотезу. То есть утверждается, что в такой регрессионной модели результат теста на мотивацию значимо лучше предсказывается средним баллом студента, чем в отсутствие этого признака.

Нужно понять, как это можно интерпретировать. Снова имеются две регрессионные модели: в первой получается, что признак не влияет значимо на отклик, а во второй — влияет, причем в отрицательную сторону. Чем меньше средний балл (рисунок 10.11), тем выше результат теста на мотивацию.

10.5.3. Разница между двумя задачами

Задачи об уровне холестерина и о мотивации отличаются тем, что признаки, которые в этих задачах используются, связаны друг с другом совершенно разными причинно-следственными конфигурациями. В первой задаче побочный признак, возраст, влияет на оба интересующих признака: и на отклик, уровень холестерина, и на признак, количество физических упражнений. Такая конфигурация называется вилкой.

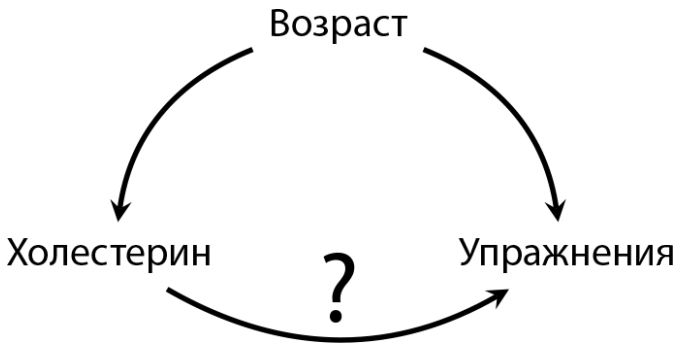

Рис. 10.12: Конфигурация ”вилка”

В задаче с поступающими конфигурация противоположная. Дело в том, что и интересующий признак, средний балл, и отклик, мотивация, наоборот, влияют на третий побочный признак, факт поступления в вуз. Такая конфигурация называется коллайдером.

Оказывается, что для того чтобы коэффициент при признаке в регрессионной модели можно было интерпретировать с точки зрения причинно-следственной связи, нужно чтобы все остальные признаки в модели были предками x, то есть влияли на x, и не были ни в коем случае потомками x, которые также одновременно являются потомками y. То есть все побочные признаки в регрессионной модели не должны быть вершинами-коллайдерами.

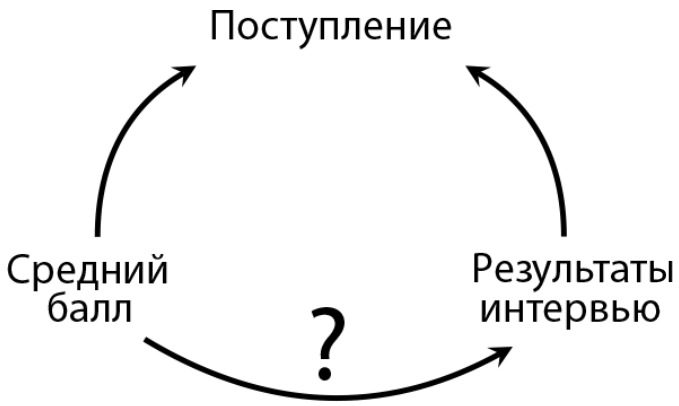

Рис. 10.13: Конфигурация "коллайдер"

Причинно-следственная связь. В линейной регрессионной модели $\hat{\beta}_1$ — это оценка среднего эффекта, то есть среднего изменения y от увеличения x_1 на 1. Этой оценке можно в некоторых случаях давать причинно-следственную интерпретацию, то есть утверждать, что если провести эксперимент, в котором зафиксированы все возможные факторы, которые могут влиять на y, и меняться будет только один из них — x_1, то именно так изменится y. Условие, при котором такую причинно-следственную интерпретацию давать можно, следующее: линейная регрессионная модель должна содержать все признаки, являющиеся причинами x_1. Кроме того, она не должна содержать признаков, которые являются следствиями одновременно x_1 и y. То есть в регрессионной модели должны быть все предки x в причинно-следственном графе и не должно быть ни одной вершины коллайдера по отношению к паре x и y.

Резюме. Итак, линейная регрессия иногда позволяет оценивать причинно-следственные связи. Однако это можно делать только при некоторых достаточно строгих предположениях: линейная модель должна быть подобрана правильно, она должна содержать правильные признаки и не содержать неправильные. Это всё надо обязательно учитывать, если требуется провести причинно-следственную интерпретацию для модели. Плохо подобранные признаки могут привести к противоположным выводам. Интересно, что на сегодняшний день существуют методы, которые по обсервационным данным позволяют восстанавливать структуру предполагаемых причинно-следственных связей между признаками в этих данных. К сожалению, эти методы достаточно сложные, и, кроме того, реализация в Python, которая существует для этих методов, находится еще в альфа-версии, поэтому эта тема в нашем курсе не рассматривается.