Практическая работа №5. Линейная регрессия. Оценка адекватности модели, оценка доверительных интервалов параметров.

Юрченков Иван Александрович, ассистент кафедры ПМ

2022-10-17

```
##
##           : 'dplyr'
##                  'package:stats':
##
##     filter, lag
##                  'package:base':
##
##     intersect, setdiff, setequal, union
## Registered S3 method overwritten by 'GGally':
##   method from
##   +.gg   ggplot2
```

Постановка задачи для выполнения практической работы

Для выполнения практического задания необходимо:

1. Открыть папку, соотвествующую своей группе.
2. Открыть папку с вариантом, совпадающим с вашим номером в списке.

В папке 3 файла с данными.

1. 1-ый файл содержит 2 ряда данных. Первый стоблец x содержит факторную переменную, второй столбец y результирующую. Для первого файла необходимо:

- Оценить коэффициент корреляции Пирсона $r(x, y)$ между двумя переменными в первом и втором столбце.
- По шкале Чеддока оценить хакактеристику корреляционной связи между величинами.
- Проверить статистическую значимость коэффициента корреляции Пирсона с помощью t-статистики.
- Построить доверительный интервал для $r(x, y)$ с надежностью $\gamma = 0.95$.
- Построить линейную регрессию между столбцами, оценить значение коэффициентов линейной зависимости.
- Оценить адекватность модели *с использованием критерия Фишера*.
- Оценить значимость полученных коэффициентов прямой.
- Построить доверительные интервалы для полученных коэффициентов.

- Оценить интервал прогноза для линейной модели на $ft = 3$ значения вперед.

2. 2-ой файл содержит 4 ряда данных. Первый ряд (столбец) содержит количественную факторную переменную, следующие два - качественную факторную переменную, последний - результирующую переменную. Для второго файла данных необходимо:

- Необходимо с помощью теста Чоу обосновать необходимость деления выборки по одной из качественных факторных переменных.
- Произвести разбиение и построить две линейных регрессии, оценить коэффициенты моделей.

3. 3-ий файл содержит 2 ряда данных. Для третьего файла данных необходимо:

- Необходимо двумя способами (тест Спирмена и тест Гольдфельда-Квандта) определить, присутствует ли в данных гетероскедастичность.
- Построить линейную регрессию, оценить значения коэффициентов модели.
- Оценить значимость полученных коэффициентов и адекватность модели.
- Все расчеты проводить для уровня значимости $\alpha = 0.05$.

Пример проведения регрессионного анализа для ряда данных

Исследуемый ряд данных

Для демонстрации проведения регрессионного анализа над рядом данных выбран набор данных цен на алмазы (diamonds), являющийся классическим набором данных для проверки регрессионных моделей и алгоритмов идентификации, очистки или корректировки выбросов. Всего в наборе данных 10 переменных. В рассмотрение возьмем только две из них:

1. carat — караты алмазов,
2. price — цена алмазов.

Предварительный анализ данных рядов показывает их нелинейную зависимость, похожую на параболическую, и чтобы избежать её в линейном регрессионном анализе, принято решение **прологарифмировать** оба ряда данных для спрямления зависимости в декартовых координатах.

Рассмотрим таблицу переменных парных данных $(ln(x), ln(y))$ одинаковой длины без пропущенных значений для данных о цене алмазов (y) с категориальными параметрами: $cut = Ideal$ (огранка), $color = J$ (цвет), $clarity = SI2$ (чистота).

Table 1: Таблица данных

n	ln(x)	ln(y)	n	ln(x)	ln(y)	n	ln(x)	ln(y)	n	ln(x)	ln(y)
1	-1.171	5.841	**31**	0.095	8.439	**61**	0.571	8.958	**91**	-1.109	5.903
2	0.020	7.965	**32**	0.231	8.446	**62**	0.531	9.048	**92**	-0.892	6.594
3	0.000	8.150	**33**	0.182	8.447	**63**	0.698	9.307	**93**	-0.942	6.111
4	0.000	8.168	**34**	0.215	8.448	**64**	0.723	9.327	**94**	-0.635	6.752
5	0.077	8.171	**35**	0.199	8.450	**65**	0.703	9.334	**95**	-0.673	6.786
6	0.049	8.193	**36**	0.239	8.454	**66**	0.723	9.336	**96**	-0.654	6.829
7	0.010	8.225	**37**	0.231	8.464	**67**	0.708	9.389	**97**	-0.616	6.886
8	0.039	8.223	**38**	0.182	8.465	**68**	0.732	9.407	**98**	-0.635	6.910
9	0.010	8.243	**39**	0.182	8.465	**69**	0.698	9.439	**99**	-0.462	6.971
10	0.058	8.227	**40**	0.239	8.472	**70**	0.718	9.446	**100**	-0.357	7.477
11	0.010	8.290	**41**	0.215	8.473	**71**	0.708	9.451	**101**	-0.357	7.510
12	0.010	8.293	**42**	0.231	8.489	**72**	0.703	9.452	**102**	-0.357	7.513
13	0.030	8.296	**43**	0.239	8.503	**73**	0.728	9.455	**103**	-0.274	7.514

n	ln(x)	ln(y)	n	ln(x)	ln(y)	n	ln(x)	ln(y)	n	ln(x)	ln(y)
14	0.030	8.307	**44**	0.239	8.521	**74**	0.698	9.458	**104**	-0.342	7.550
15	0.104	8.312	**45**	0.086	8.524	**75**	0.829	9.488	**105**	-0.329	7.563
16	0.104	8.317	**46**	0.207	8.534	**76**	0.698	9.492	**106**	-0.288	7.573
17	0.010	8.318	**47**	0.207	8.548	**77**	0.798	9.525	**107**	-0.357	7.624
18	0.104	8.319	**48**	0.293	8.570	**78**	0.829	9.527	**108**	-0.211	7.650
19	0.049	8.325	**49**	0.293	8.586	**79**	0.829	9.582	**109**	0.020	7.867
20	0.095	8.333	**50**	0.322	8.660	**80**	0.916	9.582	**110**	0.000	7.885
21	0.010	8.337	**51**	0.445	8.660	**81**	0.916	9.632			
22	0.131	8.346	**52**	0.322	8.694	**82**	0.904	9.644			
23	0.095	8.349	**53**	0.419	8.714	**83**	0.916	9.680			
24	0.113	8.372	**54**	0.315	8.715	**84**	0.928	9.680			
25	0.030	8.377	**55**	0.507	8.760	**85**	1.102	9.683			
26	0.049	8.381	**56**	0.438	8.825	**86**	0.900	9.709			
27	0.122	8.383	**57**	0.438	8.849	**87**	0.967	9.736			
28	0.174	8.414	**58**	0.464	8.876	**88**	0.959	9.753			
29	0.182	8.414	**59**	0.531	8.918	**89**	1.001	9.787			
30	0.182	8.416	**60**	0.536	8.948	**90**	0.956	9.818			

Далее наши логарифмированные данные обозначим как $x := ln(x), \quad y := ln(y),$ и примем данные переменные как рассматриваемые в нашем регрессионном анализе факторные и результирующие соответственно.

В рассматриваемой таблице данных присутствует $n = 110$ наблюдений для каждой из рассматриваемых переменных.

Построим гистограммы распределений наших данных в каждой из переменных.

Таблица гистограммы для переменной **carat** выглядит следующим образом:

```
##        groupnames abs_freq   rel_freq    low   high     med     h
## 1  (-1.17,-0.846]        4 0.03636364 -1.170 -0.846 -1.0080 0.324
## 2 (-0.846,-0.522]        5 0.04545455 -0.846 -0.522 -0.6840 0.324
## 3 (-0.522,-0.197]       10 0.09090909 -0.522 -0.197 -0.3595 0.325
## 4  (-0.197,0.128]       29 0.26363636 -0.197  0.128 -0.0345 0.325
## 5   (0.128,0.453]       28 0.25454545  0.128  0.453  0.2905 0.325
## 6   (0.453,0.777]       19 0.17272727  0.453  0.777  0.6150 0.324
## 7     (0.777,1.1]       15 0.13636364  0.777  1.100  0.9385 0.323
```

Таблица гисограммы строится для дискретной оценки непрерывного закона распределения и подсчета описательных статистик. В ней определяются из выборки следующие столбцы:

- **groupnames** — названия групп непрерывных данных, метки интервалов групп для текстового обозначения их границ нижней и верхней соответственно,
- **low** — значения нижней границы интервала данной группы,
- **med** — середина интервала группы значений,
- **high** — значения верхней границы интервала данной группы,
- **abs_freq** — значения абсолютных частот для группы значений выборки, количество значений в данной группе,
- **rel_freq** — значения относительных частот для группы значений выборки.

График гистограммы, построенной по таблице, для переменной **carat** представлен далее.

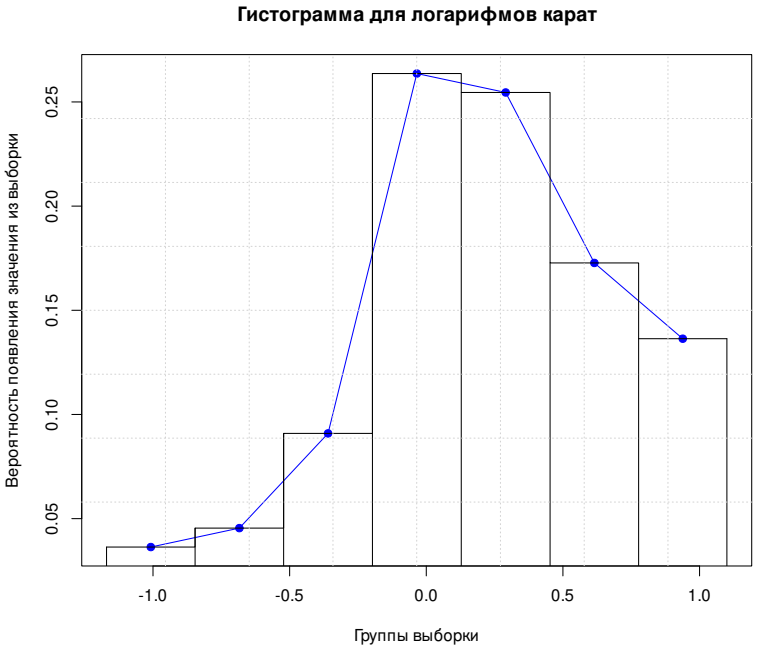
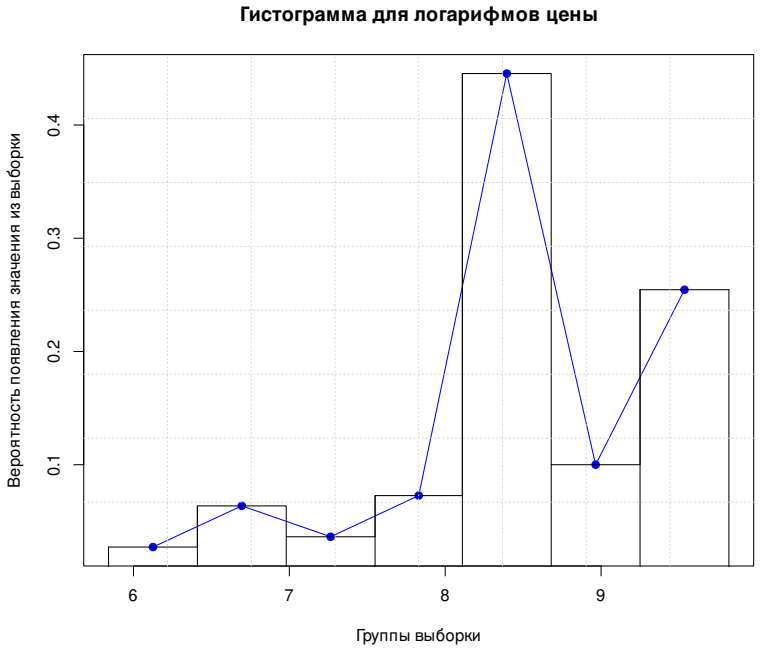


Таблица гистограммы, построенная для результирующей переменной **price** для алмазов:

```
##    groupnames abs_freq   rel_freq  low high   med    h
## 1 (5.84,6.41]        3 0.02727273 5.84 6.41 6.125 0.57
## 2 (6.41,6.98]        7 0.06363636 6.41 6.98 6.695 0.57
## 3 (6.98,7.55]        4 0.03636364 6.98 7.55 7.265 0.57
## 4 (7.55,8.11]        8 0.07272727 7.55 8.11 7.830 0.56
## 5 (8.11,8.68]       49 0.44545455 8.11 8.68 8.395 0.57
## 6 (8.68,9.25]       11 0.10000000 8.68 9.25 8.965 0.57
## 7 (9.25,9.82]       28 0.25454545 9.25 9.82 9.535 0.57
```

Также предложен график данной гистограммы для понимания присутствия в ней возможного теоретического распределения:



Для полученных выборок $X = (x_1, x_2, \dots, x_n), \quad Y = (y_1, y_2, \dots, y_n)$ наши описательные статистики рассчитываем следующим образом:

- Среднее выборочное:

$$\overline{x} = \frac{1}{n}\sum_{i=1}^{n} x_i \approx 0.219, \quad \overline{y} = \frac{1}{n}\sum_{i=1}^{n} y_i \approx 8.481$$

- Средний квадрат отклонения для гистограммы с g групп:

Для каждой из групп $i = \overline{1, g}$ подсчитаем границы Z_i и L_i:

$$g = 1 + \lfloor log_2(n) \rfloor, \quad h_x = \frac{max(X) - min(X)}{g}, \quad h_y = \frac{max(Y) - min(Y)}{g},$$

$$Z_j = min(X) + j \cdot h_x, \quad L_j = min(Y) + j \cdot h_y, \quad j = 0, 1, \dots, g.$$

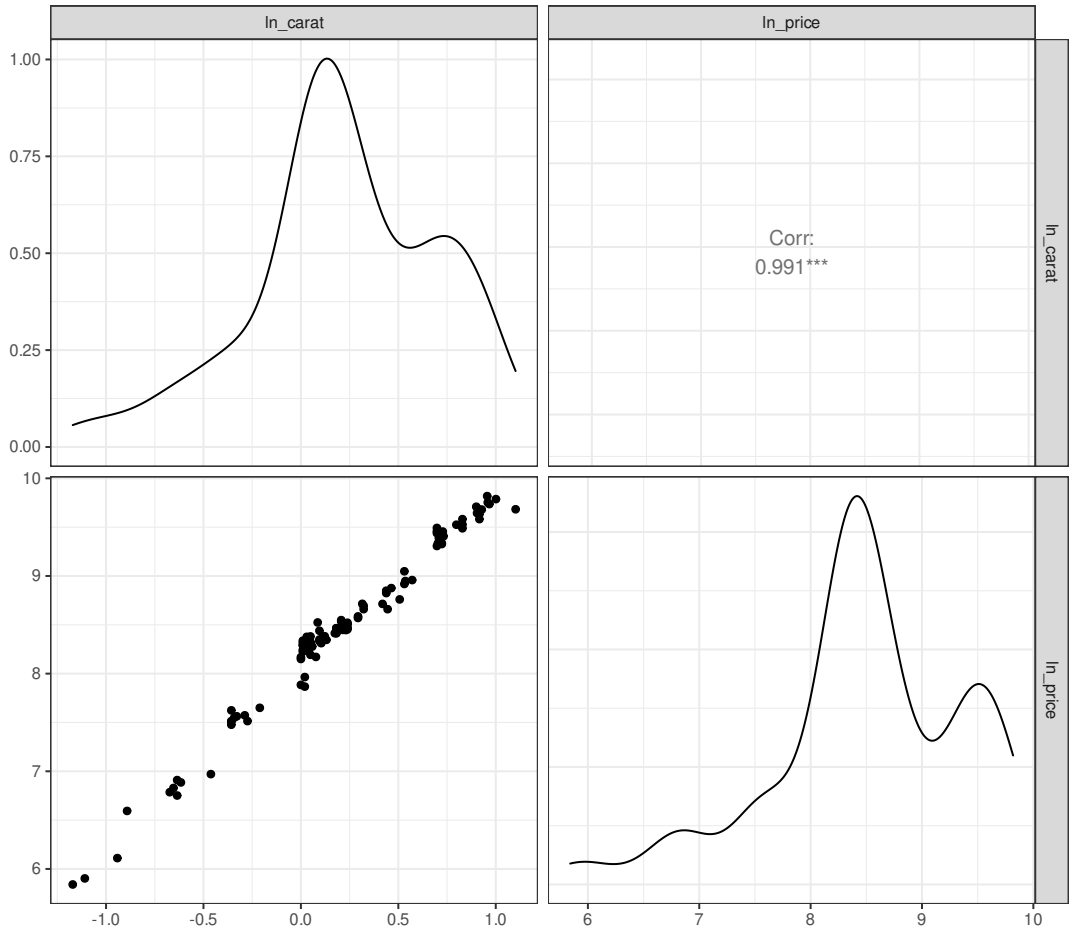
Рассчитаем середины групп ζ_i для первой и θ_i для второй:

$$\zeta_i = Z_i - Z_{i-1}, \quad \theta_i = L_i - Li - 1, \quad i = 1, 2, \dots, g.$$

Далее подсчитаем средний квадрат отклонения по определению дисперсии:

$$\sigma_x = \sqrt{\sum_{i=1}^{g} (\zeta_i - \overline{x})^2 \cdot p_i^x} \approx 0.488, \quad \sigma_y = \sqrt{\sum_{i=1}^{g} (\theta_i - \overline{x})^2 \cdot p_i^y} \approx 0.865.$$

Парный график для зависимости с отображением ядерных оценок плотности вероятности для обеих переменных:



Корреляционный анализ числовых данных

Корреляционный анализ данных позволяет ответить на вопрос о функциональной связи между двумя переменными.

Коэффициент линейной корреляции Пирсона позволяет утверждать о линейной связи между фактами (записями) переменных на основе численной оценки данной связи. Численная оценка связи между переменными определяется следующим образом:

$$r(x,y) = \frac{\overline{xy} - \overline{x} \cdot \overline{y}}{\sigma_x \sigma_y}, \quad \overline{xy} = \frac{1}{n}\sum_{i=1}^{n} x_i \cdot y_i.$$

Свойства коэффициента линейной корреляции:

1. $r(x,y) \in [-1,1]$.
2. Если $r(x,y) > 0$, то связь положительная, если $r(x,y) < 0$, то отрицательная, если $r(x,y) = 0$, то линейной связи нет.
3. Если $|r(x,y)| = 1$, то связь является линейной функциональной, то есть:

$$y = kx + b.$$

4. Чем ближе $|r(x,y)|$ к 1, тем теснее связь между исследуемыми величинами.

Рассчитаем связь между величинами по вычисленным описательным статистикам по формуле выше для $r(x, y)$. Полученное значение корреляции $r(x, y) \approx 0.973$, что говорит о сильной линейной связи между переменными. По **шкале Чеддока** данная связь характеризуется как **весьма высокая**.

Также оценим **значимость** коэффициента с помощью проверки по следующему критерию. Рассчитаем t-статистику для данного ряда данных и коэффициента линейной корреляции:

$$t_r = |r| \cdot \sqrt{\frac{n-2}{1-r^2}}.$$

Если полученное значение t-статистики выходит за границы интервала $|t_r| < t(n-2)_{1-\frac{\alpha}{2}}$, то принимается гипотеза **H_1**:

- **H_1**: значение коэффициента линейной корреляции Пирсона значительно отличается от нуля.

Если данное значение не выходит за границы, то принимается альтернативная гипотеза **H_0**:

- **H_0**: значение коэффициента линейной корреляции Пирсона незначительно отличается от нуля.

Вычислим значение t-статистики и получим, что $t_r \approx 43.479$, а значение границы интервала значимости для $t(n-2)$ распределения равно $t(n-2)_{1-\frac{0.05}{2}} = 1.98$.

Из значений t_r видно, что уровень значимости преодолен, и значение коэффициента линейной корреляции Пирсона значительно отличается от нуля и его значение является статистически значимым.

Построение линейной модели регрессии

Линейная модель регрессии — непрерывное описание линейной зависимости наблюдаемой результирующей переменной от факторной переменной:

$$\widehat{y}(x \mid a, b) = a \cdot x + b,$$

где $\widehat{y}(x)$— линейная модель зависимости результирующей переменной от факторной, a — параметр угла наклона прямолинейной зависимости в декартовой плоскости, b — параметр пересечения прямолинейной зависимости с осью $x = 0$ при нулевом значении факторной переменной.

Данные параметры модели рассчитываются исходя из решения задачи минимизации квадрата ошибок модели на пространстве параметров от имеющихся данных факторной и результирующей переменных на подвыборке :

$$||\widehat{y}(x \mid a, b) - y||^2_{L_2} \to \min_{a,b}.$$

Чаще всего решается более простая задача минимизации квадрата ошибок линейной модели по табличным данным:

$$L(a, b) = \frac{1}{n} \sum_{i=1}^{n} (a \cdot x_i + b - y_i)^2 \to \min_{a,b}.$$

Данная поставленная задача может быть решена как **аналитически**, так и численно при помощи **метода градиентного спуска**.

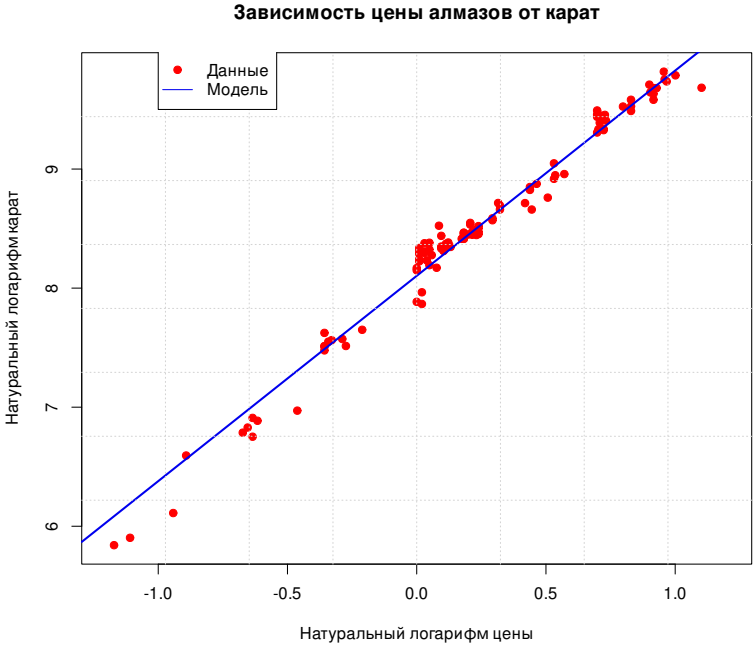
При аналитическом решении выводятся формулы для подсчета коэффициентов парной линейной регрессии и коэффициенты могут быть найдены из соотношений:

$$a = r(x, y) \cdot \frac{\sigma_y}{\sigma_x}, \quad b = \overline{y} - a \cdot \overline{x}.$$

Рассчитав коэффициенты модели по формулам получим значения:

- угол наклона прямолинейной зависимости: $a \approx 1.724$,
- пересечение с осью $x = 0$: $b \approx 8.103$.

Отобразим график парной зависимости результирующей переменной от объясняющей переменной и приведем итоговую модель зависимости:

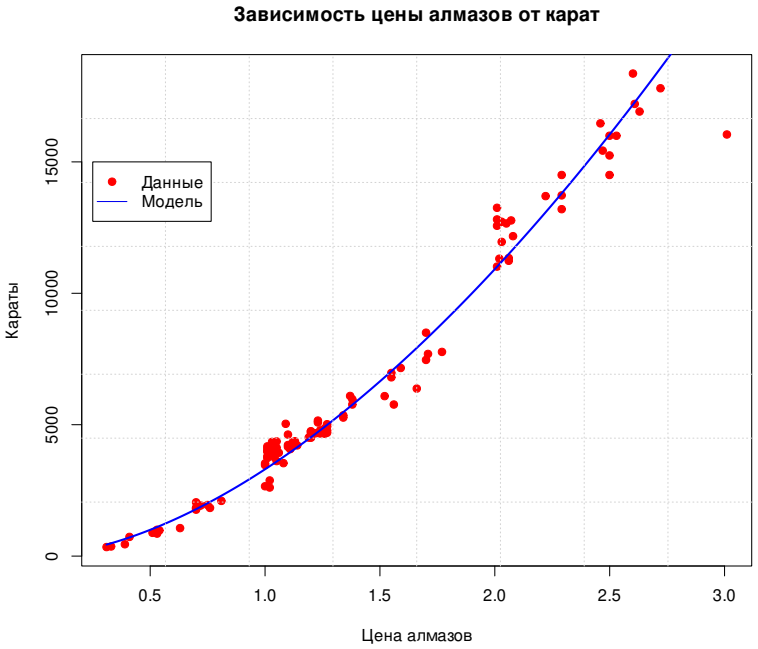


Данная зависимость имеет следующий вид:

$$ln(price) = 1.724 \cdot ln(carat) + 8.103,$$

или что в натуральном масштабе:

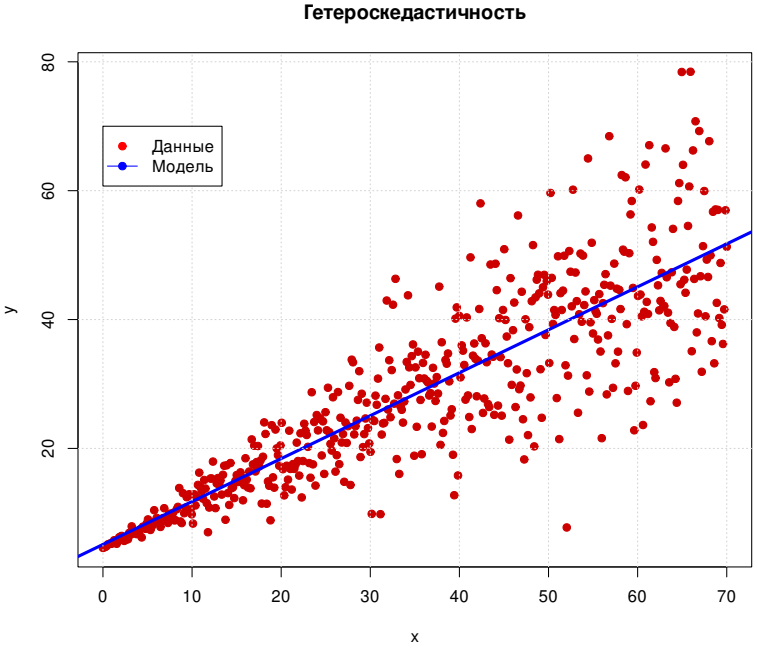
$$price = 3304.701 \cdot carat^{1.724}$$



Тест гетероскедастичности для ряда данных

Гетероскедастичность — явление неоднородности дисперсии вдоль линейной регрессионной зависимости. Данное явление сигнализирует о наличии неоднородных остатков модели регрессии и значимого отличия СКО в одном участке ошибок модели регрессии относительно данных от СКО другого участка такой метрики.

Графически гетероскедстичность выглядит следующим образом:



Хоть модель линейной регрессии и оценена через нулевое среднее ошибок, но дисперсия вдоль выборки остатков модели неоднородна и зависит от значения объясняющей переменной регрессии. В данном конкретном примере, дисперсия остатков относительно модели регрессии изменяется также линейно с ростом объясняющей переменной. Это один из возможных сценариев гетероскедастичности.

Наличие гетероскедастичности случайных ошибок приводит к неэффективности оценок, полученных с помощью метода наименьших квадратов. Кроме того, в этом случае оказывается смещённой и несостоятельной классическая оценка ковариационной матрицы МНК-оценок параметров. Следовательно, статистические выводы о качестве полученных оценок могут быть неадекватными.

Существует множество статистических тестов, позволяющих детектировать гетероскедастичность зависимости:

- тест Голдфелда — Куандта,
- тест Бройша — Пагана,
- тест Парка,
- тест Глейзера,
- тест ранговой корреляции Спирмена,
- и т.д.

В данной работе мы рассмотрим использование теста Голдфелда—Куандта для идентификации гетероскедастичности на доле данных наблюдений.

Тест Голдфелда—Куандта

Тест Голдфелда — Квандта — процедура тестирования гетероскедастичности случайных ошибок регрессионной модели, для предположения о пропорциональности случайных ошибок модели некоторой переменной (также как случай выше, самый распространенный случай).

В первую очередь, данные упорядочиваются по убыванию независимой переменной X, относительно которой имеются подозрения на гетероскедастичность.

Далее обычным МНК оценивается исходная регрессионная модель для двух разных выборок — первых m_1 и последних m_2 наблюдений в данном упорядочении, где $m_1 < n/2, \quad m_2 < n/2$. Средние $n-(m_1+m_2)$ наблюдений исключаются из рассмотрения. Чаще всего объем исключаемых средних наблюдений — порядка четверти общего объема выборки. Тест работает и без исключения средних наблюдений, но в этом случае мощность теста меньше.

Для полученных двух оценок регрессионной модели находят суммы квадратов остатков и рассчитывают F-статистику, равную отношению большей суммы квадратов остатков к меньшей

$$F = \frac{\sum\limits_{i=1}^{m_1} \left(\hat{y_1}(x_i) - y_i\right)^2 / (m_1 - k)}{\sum\limits_{i=n-m_2+1}^{n} \left(\hat{y_2}(x_i) - y_i\right)^2 / (m_2 - k)},$$

где k — число факторных (объясняющих) переменных в линейной зависимости, $\hat{y_1}(x)$ — модель на первых m_1 записях отсортированных данных по объясняющей переменной, $\hat{y_2}(x)$ — модель на последних m_2 записях отсортированных данных по объясняющей переменной.

Данный тест имеет в основе статистику, распределенную по распределению Фишера $F(m_1 - k, m_2 - k)$ с $d_1 = m_1 - k, \quad d_2 = m_2 - k$ степенями свободы. Если подсчитанная статистика по значению больше критического значения распределения Фишера с заданными степенями свободы и уровнем значимости $F(m_1 - k, m_2 - k)_{1-\frac{\alpha}{2}}$, то нулевая гипотеза отвергается и **гетероскедастичность имеет место** для заданной линейной зависимости.

Тест на гетероскедастичность для данных цен алмазов с использованием теста Голдфелда—Куандта

Отсортируем выборку по переменной $carat$ и возьмем по $m_1 = m_2 = \lfloor 3n/8 \rfloor$ записей с каждой стороны, что в сумме составит $m_1 + m_2 \approx \lfloor 3n/4 \rfloor$ записей. Тогда можем рассчитать F-статистику:

Для $n = 110$, $m_1 = m_2 = 41$, $k = 1$ ввиду того что мы имеем дело с парной регрессией на одну объясняющую (факторную переменную).

Модели на подвыборках оценены как:

- $\widehat{y_1}(x) = 2.03 \cdot x + 8.17$,
- $\widehat{y_2}(x) = 1.719 \cdot x + 8.102$.

Тогда сумма квадратов ошибок для первой и второй модели соотвественно равны $RSS_1 \approx 0.669, \quad RSS_2 \approx 0.415$ и имеем расчетную статистику:

$$F = \frac{0.669/40}{0.415/40} = \frac{0.669}{0.415} \approx 1.6108.$$

Получаем расчетную статистику $F \approx 1.6108$, при критическом значении $F(40,40)_{1-\frac{0.05}{2}} \approx 1.875$. Рассчетная статистика ненамного меньше критического значения распределения Фишера, что всё же говорит о принятии нулевой гипотезы об отсутствии гетероскедастичности в данной зависимости.

В полученной ситуации требуются повторные уточняющие тесты с различными значениями m_1 и m_2 для достаточной достоверности теста.

Оценка адекватности оцененной модели регрессии

Проверка значимости полученной модели называется проверкой адекватности. Одним из способов проверки значимости линейной модели регрессии является использование критерия Фишера, который заключается в расчёте $F(n-2, n-1)$-распределенной статистике, определенной как:

$$F = \frac{S^2_{\text{АД}}}{S^2_{\text{ОБЩ}}}, \quad S^2_{\text{АД}} = \frac{\sum\limits_{i=1}^{n} \left(y_i - \widehat{y}(x_i)\right)^2}{n-2}, \quad S^2_{\text{ОБЩ}} = \frac{\sum\limits_{i=1}^{n} \left(y_i - \overline{y}\right)^2}{n-1}.$$

Если полученное значение F-статистики равно или выше критического значения $F(n-2, n-1)_{1-\frac{\alpha}{2}}$ при заданном уровне значимости α, то модель признается неадекватной и принимается гипотеза H_1, в альтернативном случае принимается гипотеза H_0 и модель признается адекватной.

Рассмотрим рассчет статистики на примере данных о цене алмазов.

Расчет статистики для оценки адекватности модели регрессии цен на алмазы

Для полученной линейной модели зависимости натурального логарифма цены алмазов от натурального логарифма карат произведем расчет полной суммы квадратов (TSS) и суммы квадратов ошибок (RSS) как составных частей F-статистики:

$$RSS = \sum_{i=1}^{n} (y_i - \widehat{y}(x_i))^2 \approx 1.648, \quad TSS = \sum_{i=1}^{n} (y_i - \overline{y})^2 \approx 82.185.$$

Теперь произведем расчет F-статистики и получим:

$$F = \frac{RSS/(n-2)}{TSS/(n-1)} = \frac{1.648 \cdot 109}{82.185 \cdot 108} \approx 0.0202$$

Критическое значение при уровне значимости $\alpha = 0.05$ равно $F(108,109)_{1-\frac{0.05}{2}} = 1.459577$. Рассчетная F-статистика значительно меньше критического значения распределения Фишера, что говорит об адекватности полученной линейной модели регрессии.

Оценка статистической значимости коэффициентов линейной модели регрессии

Оценка прогнозного интервала для линейной модели регрессии

Темы вопросов на защиту практической работы

1. Задачи корреляционного анализа. Выборочный коэффициент линейной корреляции (Пирсона) и его свойства. Шкала Чеддока.
2. Выборочный коэффициент линейной корреляции (Пирсона) и его свойства. Оценка значимости коэффициента корреляции.
3. Корреляция и причинная связь. Проблемы корреляционного анализа.
4. Ранговая корреляция. Коэффициент ранговой корреляции Спирмена.
5. Задачи регрессионного анализа. Функциональная и статистическая связь. Аппроксимационные модели. Параметрическое множество функций.
6. Линейная регрессия. Определение коэффициентов линейной модели методом наименьших квадратов.
7. Проверка значимости полученных коэффициентов модели. Проверка адекватности модели с помощью критерия Фишера.
8. Доверительный интервал прогноза. Проверка адекватности модели с помощью критерия Фишера.