Линейная регрессия. Оценка адекватности модели, оценка доверительных интервалов параметров

Постановка задачи

Скачать папку с исходными данными. Открыть папку, соответствующую номеру своей группы. Открыть папку, соответствующую номеру своего варианта. В папке data можете найти 3 файла.

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

1. Оценить коэффициент корреляции Пирсона $r(x, y)$ между двумя переменными в первом и втором столбце

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 с помощью коэффициента корреляции Пирсона рассчитывается по следующей формуле:

$$r = \frac{\overline{xy} - \overline{x}\,\overline{y}}{S_x S_y}$$

Выборочные средние

$$\overline{x} = \frac{1}{n}\sum x_i \qquad \overline{y} = \frac{1}{n}\sum y_i \qquad \overline{xy} = \frac{1}{n}\sum x_i y_i$$

$$S_x = \sqrt{\frac{1}{n}\sum {x_i}^2 - \bar{x}^2} \qquad S_y = \sqrt{\frac{1}{n}\sum {y_i}^2 - \bar{y}^2}$$

Или

$$r = \frac{\sum_{i=1}^{n}(x_i - \bar{x})(y_i - \bar{y})}{\sqrt{\sum_{i=1}^{n}(x_i - \bar{x})^2 \sum_{i=1}^{n}(y_i - \bar{y})^2}}$$

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 связь отрицательная, если $r(x,y) = 0$, то линейной связи нет.

3. Если $|r(x,y)| = 1$, то связь является линейной функциональной, то есть $y = ax + b$

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

2. По шкале Чеддока оценить характеристику корреляционной связи между величинами

Шкала Чеддока

Диапазон изменения $\|r\|$	0.1–0.3	0.3–0.5	0.5–0.7	0.7–0.9	0.9–0.99
Характеристика связи	Слабая	Умеренная	Заметная	Высокая	Весьма высокая

3. Проверить статистическую значимость коэффициента корреляции Пирсона с помощью t-статистики

Значимость коэффициента корреляции оценивается с помощью t-статистки

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

$\mathbf{H_0}$: r $= 0$. Коэффициент корреляции не отличается от нуля.

$\mathbf{H_1}$: r $\neq 0$. Коэффициент корреляции значительно отличается от нуля.

Если $|t| > t_{1-\frac{\alpha}{2};n-2}$, то гипотеза H_0 отвергается.

4. Построить линейную регрессию между столбцами, оценить значение коэффициентов линейной зависимости

Значения коэффициентов линейной регрессии находится из решения задачи минимизации квадрата ошибок модели на пространстве параметров от имеющихся данных факторной и результирующей переменных на выборке:

$$J(a,b) = \sum_{i=1}^{n} \left(y_i - (ax_i + b)\right)^2 \to min$$

Данная задача может быть решена численно или аналитически. Аналитическое решение находится путем приравнивания к нулю производных от функции $J(a,b)$ по параметрам a и b:

$$\begin{cases} \frac{\partial J}{\partial a} = 0 \\ \frac{\partial J}{\partial b} = 0 \end{cases}$$

$$\begin{cases} 2 \cdot \sum_{i=1}^{n} (y_i - ax_i - b) \cdot (-x_i) = 0 \\ 2 \cdot \sum_{i=1}^{n} (y_i - ax_i - b) \cdot (-1) = 0 \end{cases}$$

$$\begin{cases} \sum_{i=1}^{n} y_i - \sum_{i=1}^{n} ax_i - \sum_{i=1}^{n} b = 0 \\ \sum_{i=1}^{n} y_i x_i - \sum_{i=1}^{n} ax_i{}^2 - \sum_{i=1}^{n} bx_i = 0 \end{cases}$$

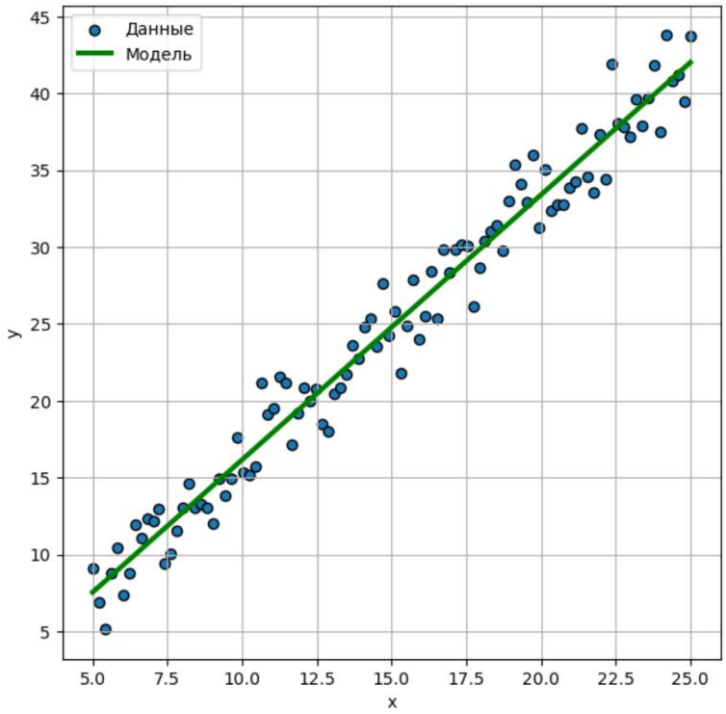
$$\begin{cases} \sum_{i=1}^{n} y_i = a \sum_{i=1}^{n} x_i + nb \\ \sum_{i=1}^{n} y_i x_i = a \sum_{i=1}^{n} x_i{}^2 + b \sum_{i=1}^{n} x_i \end{cases}$$

$$\begin{cases} a = \frac{\sum_{i=1}^{n} y_i \sum_{i=1}^{n} x_i{}^2 - \sum_{i=1}^{n} x_i \sum_{i=1}^{n} y_i x_i}{n \sum_{i=1}^{n} x_i{}^2 - (\sum_{i=1}^{n} x_i)} \\ b = \frac{n \sum_{i=1}^{n} y_i x_i - \sum_{i=1}^{n} x_i \sum_{i=1}^{n} y_i}{n \sum_{i=1}^{n} x_i{}^2 - (\sum_{i=1}^{n} x_i)} \end{cases}$$

или

$$\begin{cases} a = r \cdot \dfrac{\sigma_y}{\sigma_x} \\ b = \dfrac{1}{n}\sum y_i - r \cdot \dfrac{\sigma_y}{\sigma_x} \cdot \dfrac{1}{n}\sum x_i \end{cases}$$

После нахождения коэффициентов прямой необходимо отобразить график парной зависимости результирующей переменной от объясняющей переменной и привести итоговую модель зависимости.



5. Оценить адекватность модели с использованием критерия Фишера

Проверка значимости полученной модели называется проверкой адекватности. Хотим узнать, полученная линейная модель по параметрам удовлетворительно описывает экспериментальные данные или нет.

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:

$$F = \frac{S_Б^2}{S_M^2} \qquad S_M^2 = \min\left(S_{мод}^2, S_{общ}^2\right) \qquad S_Б^2 = \max\left(S_{мод}^2, S_{общ}^2\right)$$

$$S_{мод}^2 = \frac{\sum\left(y_i - \hat{y}(x_i)\right)^2}{n - d_\theta}$$

$d_\theta -$ размерность вектора параметров

Дисперсия адекватности $S^2_{\text{мод}}$ характеризует величину среднего разброса экспериментальных точек относительно линии регрессии. Она позволяет оценить ошибку, с которой уравнение регрессии предсказывает фактический результат. Минимальная величина остаточной дисперсии должна свидетельствовать о более удачном выборе линии регрессии.

$$S^2_{\text{общ}} = \frac{\sum(y_i - \bar{y})^2}{n-1}$$

$S^2_{\text{общ}}$ – это усредненная, или общая дисперсия. В качестве таковой принимается квадрат стандартной ошибки. Этот показатель фактически характеризует случайную ошибку для всей выборки, т. е. оценивает несоответствие между конкретными (текущими) значениями результата эксперимента и средним арифметическим.

Если полученное значение F-статистики равно или выше критического значения $F\ (a, n-2, n-1)$.

$\mathbf{H_0}$: Модель не адекватна. Дисперсии равны.

$\mathbf{H_1}$: Модель адекватна. Дисперсии не равны.

Если $F > F_{\alpha;n-2,n-1}$, то гипотеза H_0 отвергается. Дисперсии не равны, а значит модель признается адекватной, она удовлетворительно описывает экспериментальные данные с заданной степенью достоверности (надежности).

6. Оценить значимость полученных коэффициентов линейной регрессии

Наряду с общей проверкой модели можно так же проверить значимость каждого коэффициента. В основе проверки лежит гипотеза о равенстве параметров нулю. Для линейной регрессии считаются следующие показатели:

$$m_a = \frac{S_{\text{мод}}}{\sigma_x \sqrt{n}}$$

$$m_b = \frac{S_{\text{мод}} \sqrt{\sum x_i^2}}{\sigma_x n}$$

Тогда предложенные ниже статистики распределены по t-распределению с $df = n - 2$ степенями свободы.

$$T_a = \frac{a}{m_a}$$

$$T_b = \frac{b}{m_b}$$

Критические значения для полученных статистик определяются по таблице t-распределения $t\left(1 - \frac{a}{2}; n - 2\right)$.

Если $|T_a| > t\left(1 - \frac{a}{2}; n - 2\right)$, $|T_b| > t\left(1 - \frac{a}{2}; n - 2\right)$, то параметр признается значимым.

7. Построить доверительные интервалы для полученных коэффициентов

Доверительные интервалы параметров рассчитываются для каждого из параметров.

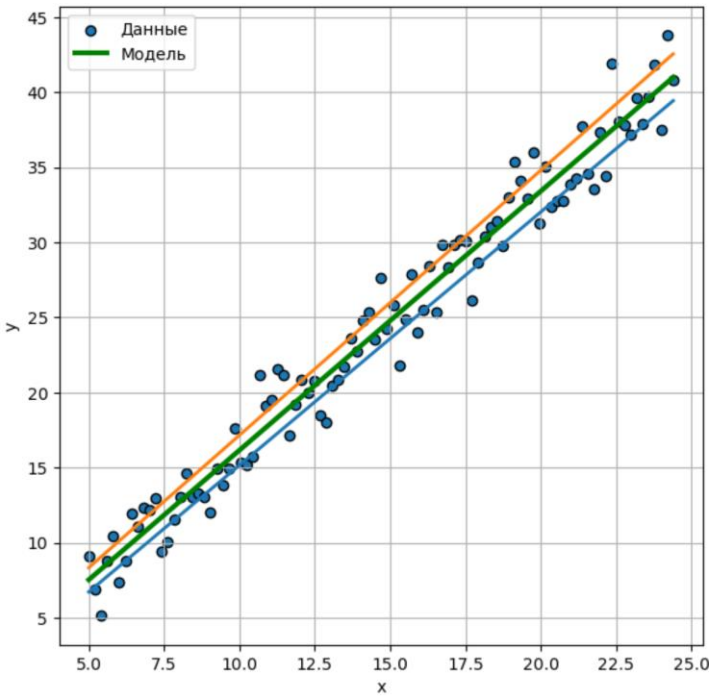
Доверительный интервал для параметра a

$$a - t\left(1 - \frac{\alpha}{2}; n - 2\right) m_a < a < a + t\left(1 - \frac{\alpha}{2}; n - 2\right) m_a$$

Доверительный интервал для параметра b

$$b - t\left(1 - \frac{\alpha}{2}; n - 2\right) m_b < b < b + t\left(1 - \frac{\alpha}{2}; n - 2\right) m_b$$

Изобразить модели, отвечающие граничным значениям полученных интервалов.



8. Оценить интервал прогноза для линейной модели на $\Delta x = 3$ значения вперед

Интервал предсказания (доверительный интервал прогноза) – интервал, который с заданной вероятностью (1 – α) содержит фактическое значение целевой переменной y. Если прогнозное значение факторного признака x подставить в уравнение регрессии, мы получаем для него предсказанное значение $\hat{y}$,которое является точечной оценкой для фактического y.

Интервальные оценки прогноза модели регрессии необходимы для оценки верхнего и нижнего предела значений, возможных при применении модели регрессии для экстраполяции данных. Данные интервалы являются более важными самого прогноза, так как являются инструментом понижения рисков при прогнозах важных показателей в будущее.

Прогнозный интервал модели регрессии является схожим инструментом с оценкой доверительных интервалов модели и отвечает на вопрос об интервале будущих значений показателя с заданным уровнем значимости α. Чем ниже его значение, тем больше будет уровень надежности $\gamma = 1 - \alpha$, тем шире будет прогнозный интервал и менее информативным будет прогноз.

Для начала необходимо определиться со значением x для которого нам хочется получить прогнозное значение $\hat{y}$ по модели $\hat{y}(x)$, подставив в выражение

для модели данное значение. Получившееся значение будет лежать в интервале с уровнем надежности γ.

Интервальные оценки для прогноза линейной модели регрессии рассчитываются следующим образом:

$$\hat{y} \in [\hat{y}(x) - E; \hat{y}(x) + E]$$

$$E = t(1 - \frac{\alpha}{2}; n - 2) \cdot S_{\text{мод}} \sqrt{1 + \frac{1}{n} + \frac{(x - \bar{x})^2}{n\sigma_x^2}}$$

Полученные верхняя и нижняя оценка данного интервала называется интервальной оценкой прогноза.

Второй 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:

1. С помощью теста Чоу обосновать необходимость деления выборки по одной из качественных факторных переменных. Произвести разбиение и построить две линейные регрессии, оценить коэффициенты моделей.

Пусть есть выборка, имеющая один факторный признак x, целевой признак y и один бинарный номинальный признак (переменную) d. К такой выборке всегда можно привести любую выборку применяя процедуру создания фиктивных переменных из категориальных.

Для такой выборки ставится вопрос о её делении и обучении двух моделей на разных подвыборках (x_1, y_1, d_1), $d_1 = 0$ и (x_2, y_2, d_2), $d_2 = 1$:

$$\begin{cases} \hat{y}_1(x) = a_1 x + b_1, & \forall x: d = 0 \\ \hat{y}_2(x) = a_2 x + b_2, & \forall x: d = 1 \end{cases}$$

или обучении одной модели на всех данных выборки (x, y, d):

$$\hat{y}(x) = ax + b$$

Один из подходов к определению важных для модели качественных переменных состоит в проверке того, насколько однородными будут подвыборки после деления основной выборки по номинальному признаку.

Тест Чоу позволяет сказать, есть ли смысл учитывать номинальный признак в модели.

Для всех моделей в тесте считается RSS (сумма квадратов остатков модели).

Для первой модели − сумма остатков по данным, которые доступны только ей, для второй модели аналогично, для общей модели – сумма квадратов остатков по всем данным. Далее рассчитывается F-статистика по формуле:

$$F = \frac{(RSS - RSS_1 - RSS_2)/k}{(RSS_1 + RSS_2)/(n-2k)},$$

где RSS_1 и RSS_2 – сумма квадратов остатков первой ($d = 0$) и второй ($d = 1$) модели соответственно, RSS – сумма квадратов остатков стандартной модели на всей выборке, а k – количество параметров линейной модели регрессии (для стандартной парной зависимости $k = 2$).

Эта статистика имеет распределение Фишера $F \sim F(k, n-2k)_\alpha$

$$RSS_m = \sum_{i=1}^{n} \left(y_i - \hat{y}_m(x_i)\right)^2$$

$$k - \text{размерность вектора параметров}$$

Нулевая гипотеза – параметры моделей равны, то есть деление по выбранному номинальному признаку не имеет смысла (разделение выборки на две подвыборки не приводит к улучшению качества модели)

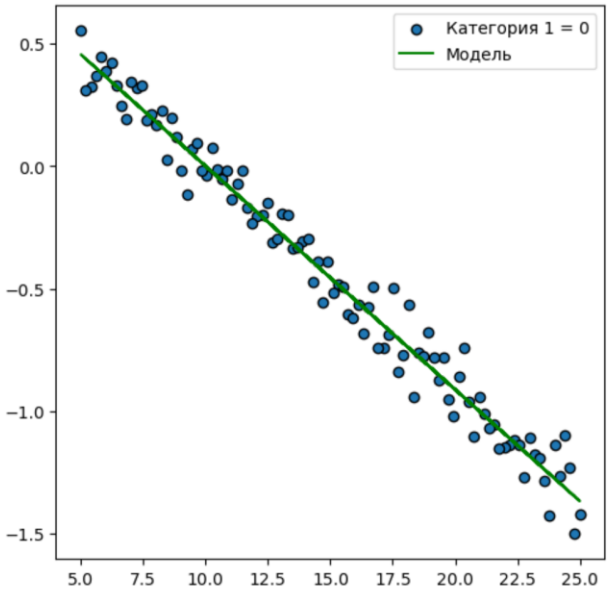
$$H_0: a_1 = a_2 \text{ и } b_1 = b_2$$

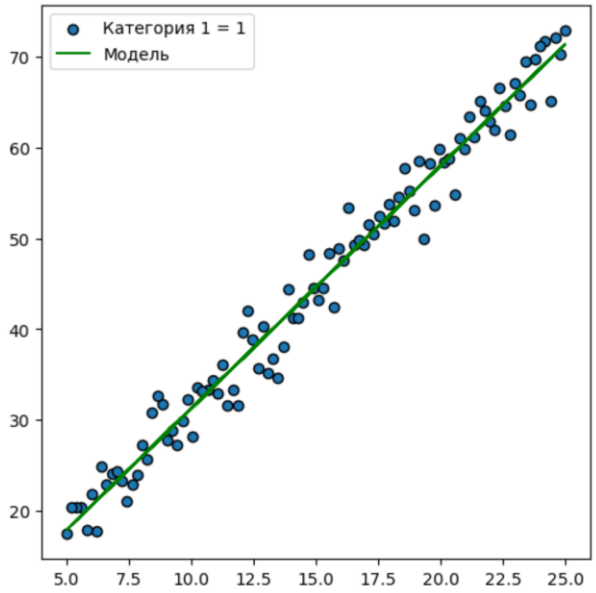
Альтернативная гипотеза – параметры моделей не равны, то есть деление по выбранному номинальному признаку имеет смысл. (Разделение выборки на две подвыборки приводит к улучшению качества модели)

$$H_1: a_1 \neq a_2 \text{ или } b_1 \neq b_2$$

Если полученный критерий выше критического значения, то нулевая гипотеза отвергается, а значит подвыборки неоднородны и деление по выбранному признаку имеет смысл.

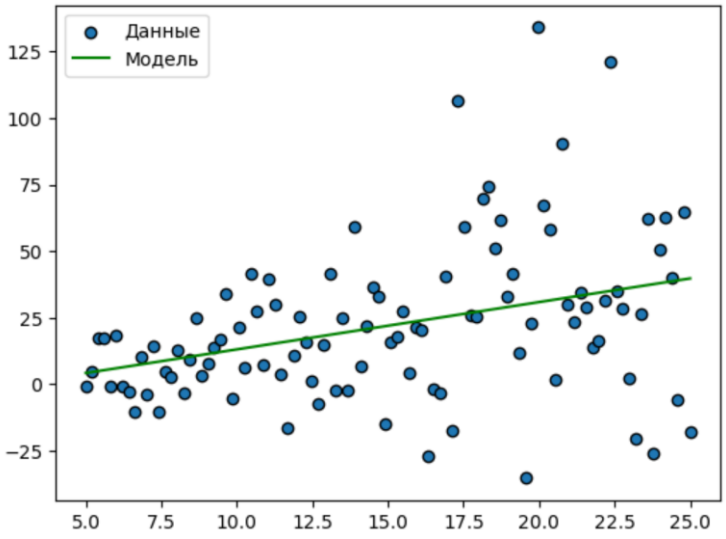
Необходимо провести тест Чоу для каждого номинального признака и построить графики линейной регрессии.





Третий файл содержит 2 ряда данных. Для третьего файла необходимо:

1. Построить линейную регрессию, оценить значения коэффициентов модели.



$$a = 1.78 \qquad b = -4.72$$

$$\hat{y}(x) = 1.78x - 4.72$$

2. Оценить значимость полученных коэффициентов и адекватность модели.

3. Двумя способами (тест Спирмена и тест Гольдфельда-Квандта) определить, присутствует ли в данных гетероскедастичность.

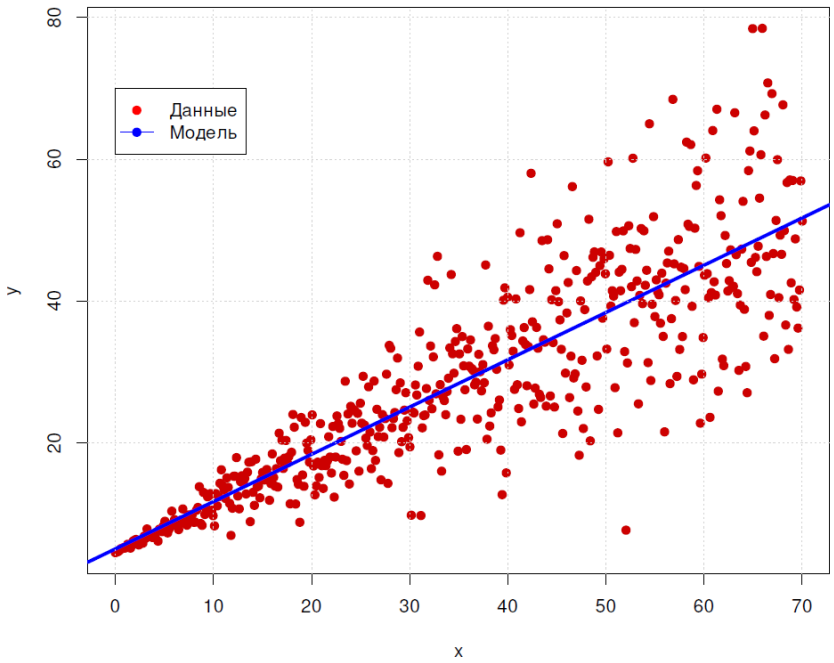
Определение гетероскедастичности

Гетероскедастичность – явление неоднородности дисперсии вдоль линейной регрессионной зависимости.

Гомоскедостичность – явление однородности дисперсии вдоль линейной регрессионной зависимости.

Данное явление сигнализирует о наличии неоднородных остатков модели регрессии. Значимого отличия стандартного отклонения на одном участке ошибок модели регрессии относительно данных от стандартного отклонения другого участка.

Графически гетероскедастичность выглядит следующим образом:



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.

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.

Тест ранговой корреляции Спирмена для идентификации гетероскедастичности

Предполагается, что дисперсия остатков будет увеличиваться или уменьшаться по мере увеличения x.

Сначала нужно рассчитать остатки модели:

$$e_i = |y_i - \hat{y}(x_i)|$$

Далее необходимо ранжировать x и e. Отсортировав отдельно обе выборки по возрастанию и присвоив каждому значению выборки ранг позиции в отсортированной выборке. Пример новых переменных:

	x	y	e	rangX	rangE
0	5.000000	4.563998	7.525855	1	4
1	5.202020	9.296003	2.971615	2	2
2	5.404040	11.964633	0.480749	3	1
3	5.606061	17.691032	5.067886	4	3
4	5.808081	4.680889	8.120022	5	5

Значения переменных rank – это позиции элементов в отсортированной выборке. Для похожих значений выбирается среднее их рангов.

Далее для ранжированных данных ошибок e_i и факторной переменной x_i производится подсчёт коэффициента ранговой корреляции Спирмена:

$$r_s(e,x) = 1 - \frac{6\sum_{i=1}^{n}(rank(e_i) - rank(x_i))^2}{n(n^2-1)}$$

Статистика рассчитывается по следующей формуле:

$$t_s = \frac{r_s\sqrt{n-2}}{\sqrt{1-{r_s}^2}}$$

Если $t_s > t_{\text{кр}}(\text{n}-2)$, то гетероскедастичность присутствует. Дисперсия вдоль линейной регрессионной зависимости неоднородна.

Тест Гольдфельда-Квандта

Тест Голдфелда-Квандта — процедура тестирования гетероскедастичности ошибок регрессионной модели для предположения о пропорциональности случайных ошибок модели некоторой переменной.

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

Далее оценивается исходная регрессионная модель для двух разных выборок — первых m_1 и последних m_2 наблюдений в данном упорядочении, где $m_1 < \frac{n}{2}$, $m_2 < \frac{n}{2}$. Рекомендуется брать $m = \frac{3}{8}n$. 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 критерия меньше. Если имеет место гетероскедастичность и если предположение относительно ее природы верно, то дисперсия в последних m_2 наблюдениях будет больше, чем в первых m_1, и это будет отражено в сумме квадратов остатков в частных регрессиях.

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

Для сравнения дисперсий двух выборок используется тест Фишера:

$$F = \frac{\frac{\sum_{i=1}^{m_1}(\widehat{y_1}(x_i) - y_i)^2}{m_1 - 1}}{\frac{\sum_{i=n-m_2+1}^{n}(\widehat{y_2}(x_i) - y_i)^2}{m_2 - 1}}$$

$$F \sim F_{m_1-1;m_2-1}$$

Данные тест имеет в основе статистику, распределенную по распределению Фишера с $d_1 = m_1 - 1$ и $d_2 = m_2 - 1$ степенями свободы.

Если $F > F_{\text{кр}}$, то дисперсии не равны, а значит присутствует гетероскедастичность для заданной линейной зависимости.

Если $F < F_{\text{кр}}$, то дисперсии равны, а значит нет гетероскедастичности.

Все расчеты проводить для уровня значимости α = 0.05.

Структура отчета

5 ЛИНЕЙНАЯ РЕГРЕССИЯ. ОЦЕНКА АДЕКВАТНОСТИ МОДЕЛИ, ОЦЕНКА ДОВЕРИТЕЛЬНЫХ ИНТЕРВАЛОВ ПАРАМЕТРОВ

5.1 Постановка задачи

5.2 Ход выполнения работы

5.2.1 Оценка коэффициента корреляции Пирсона. Оценка характеристики корреляционной связи по шкале Чеддока

5.2.2 Проверка статистической значимости коэффициента Корреляции

5.2.3 Построение модели линейной регрессии

5.2.4 Оценка адекватности модели

5.2.5 Оценка значимости коэффициентов модели

5.2.6 Построение доверительных интервалов коэффициентов модели

5.2.7 Оценка интервала прогноза линейной модели

5.2.8 Проведение теста Чоу

5.2.9 Построение модели линейной регрессии. Оценка значимости коэффициентов и адекватности модели

5.2.10 Проверка данных на гетероскедастичность при помощи теста Гольдфельда-Квандта и теста Спирмена

5.3 Вывод