Эконометрика

ПМ-1701

Преподаватель:

Курышева Светлана Владимировна

Санкт-Петербург
2020 г., 6 семестр

Список литературы

[1] И.И.Елисеева, Эконометрика

[2] И.И.Елисеева, Практикум по эконометрике

[3] Мхитаран, Эконометрика

[4] Доугерти, Введение в эконометрику

[5] Брендт Э., Практика эконометрики

Содержание

1 07.02.2020

1.1 Общие понятия об эконометрике

Эконометрика - это наука, которая дает конкретное количественное выражение закономерностям и взаимосвязям экономических явлений и процессов с помощью статистико-математических методов и моделей.

Связь эконометрики с другими науками:

- Экономическая теория (сущность связи явлений)
- Статистика (информационная база)
- Математические и статистические методы:
  - $C = k \cdot Y + L$, $0 < |k| < 1$ - регрессия
  - $r = sC + Dx + T$, где t - сбережения, а x - инвестиции
  - Если s=D, то $t = s(C + x) + T$ - уравнение двухфакторной регрессии
  - $Y + r = C + x$ - балансовое тождество

Этапы построения эконометрической модели:

1. Теоретоическое описание рассмариваемого процесса
2. Сбор данных, анализ их качества
3. Спецификация модели
   (a) Выявление объясняемых (Y) и объясняющих (X) переменных
   (b) Выбор функций
4. Оценка параметров модели
5. Верификация модели (т.е проверка достоверности)
6. Интерпертация результатов

1.2 Парная регрессия и корреляция в эконометрических исследованиях

Последовательность анализа регрессии:

1. Выбор типа математической функции при построении уравнения регрессии
2. Оценка параметров уравнения

3. Показатели силы связи

4. Статистическая оценка достоверности (F-критерий Фишера)

5. Интервальная оценка параметров уравнений парной регрессии

6. Использование модели

Выбор функции для модели может проводиться 3-мя способами

1. Аналитический

2. Графический

3. Экспериментальный

Основные виды функций в модели парной регресии:
$y = ax + b, y = a + \frac{b}{x}, y = a + bx + cx^2, y = ax^b, a = b^x, y = ae^{bx}$

1.3 Предпосылки регрессионной модели

1. Модель линейна по параметрам

2. $\mathbb{E}\xi_i = 0 \forall i$, т.е ожидание значения случайного члена должно быть равно нулю в каждом наблюдении из-за того, что каждое наблюдение не должно включать в себя смещения ни в каком из направлений.

3. $\mathbb{D}\xi_i = Const$, т.е его значение в каждом наблюдении получено из распределения с постоянной теоретической дисперсией. Также не должно быть причин , делающих его больше подверженным ошибке в одних наблюдениях по сравнению с другим. Заметим, что

$$\mathbb{E}\xi_i^2 = \mathbb{D}\xi_i = \mathbb{D}\sigma^2_{\xi_i} | \forall i$$

4. Значения случайного члена имеют взаимно независимые распределения. Случайный член не подвержен автокорреляции, т.е отсутствует систематическая связь между его значениями в любых двух наблюдениях. Ковариация равна нулю:

$$\sigma_{\xi_i \xi_j} = \mathbb{E}(\xi_i \xi_j) = \mathbb{E}\xi_i \cdot \mathbb{E}\xi_j = 0 | \forall i \neq j$$

5. $\xi_i \sim \mathbb{N}(0, \sigma^2)$: если случайный член нормально распределен, то распределены нормально и коэффициенты регрессии.

1.4 Оценка параметров модели

Рассмотрим случаи, для которых мы хотим предпооложить, что одна *зависимая* переменная Y определяется другими переменными, называемые *объясняющими* переменными (регрессорами). Математическая зависимость, связывающая эти переменные, называется *моделью регрессии.* Мы допускаем, что модель регрессии имеет факт неточности - *случайный* (остаточный) член.

Начнем с рассмотрения простешей модели:

$$Y_i = \alpha + \beta X_i + \xi_i \tag{1}$$

Y_i - значение зависимой переменной, α и β - постоянные величины - параметры уравнения, ξ_i - случайный член.

Задача регрессионого анализа состоит в получении оценок α и β и, следовательно, в определении положения прямой по точкам $\Leftrightarrow$ нужно посроить прямую, в наибольшей степени соответствующую этим точкам.

a - отсечение Y - оценка α

b - угловой коэффициент - оценка β

Пусть

$$\hat{Y} = a + bX_i \tag{2}$$

оцениваемая модель, а Y_i - оцененное значение Y. Наша задача заключается в том, чтобы выяснить, существут ли способы оценки коэффициентов a, b алгебраическим путем.

Обозначим за

$$e_i = Y_i - \hat{Y}_i = Y_i - a - bX_i \tag{3}$$

Остаток наблюдений зависит от выбора коэффициентов a и $b \Rightarrow$ задача заключается в том, чтобы выбрать такие a и b, предсказанное значение функции от искомой в каждой точке было минимальным. Глупо минимизировать сумма остатков, потому что при выборе выборочного среднего модели:

$$\sum e_i = 0 \tag{4}$$

Поэтому будем минимизировать сумму квадратов остатков. Данный метод называется *Методом Наименьших Квадратов* или сокращенно МНК.

1.4.1 Метод наименьших квадратов

Пусть у нас имеются n наблюдений (X_i, Y_i), Y зависит от X и мы хотим подобрать уравнение:

$$\hat{Y} = a + bX_i$$

Запишем формально нашу задачу в обозначениях метода наименьших квадратов (МНК):

$$S = \sum (Y_i - \hat{Y}_i)^2 = \sum (Y_i - a - bX_i)^2 \to min \tag{5}$$

$$\begin{cases} \frac{\partial S}{\partial a} = -2\sum Y + 2na + 2b\sum X = 0 \\ \frac{\partial S}{\partial b} = -2\sum YX + 2a\sum X + 2b\sum X^2 = 0 \end{cases} \tag{6}$$

Применение метода наименьших квадрато приводит к системе уравнений, которая для линейных уравнений имеет вид:

$$\begin{cases} \sum Y = na + b\sum X \\ \sum YX = a\sum X + b\sum X^2 \end{cases} \tag{7}$$

Решим данную систему линенйных уравнений методом Крамера:

$$\delta = \begin{vmatrix} n & \sum X \\ \sum X & \sum X^2 \end{vmatrix} = n \cdot \sum X^2 - (\sum X)^2$$

$$\delta_b = \begin{vmatrix} n & \sum Y \\ \sum X & \sum XY \end{vmatrix} = n \cdot \sum XY - \sum X \sum Y$$

$$b = \frac{\delta_b}{\delta} = \frac{n \cdot \sum XY - \sum X \sum Y}{n \cdot \sum X^2 - (\sum X)^2} = \frac{\frac{\sum XY}{n} - \frac{\sum X \sum Y}{n}}{\frac{\sum X^2}{n} - \frac{(\sum X)^2}{n^2}} = \frac{\overline{XY} - \overline{X} \cdot \overline{Y}}{\overline{X^2} - (\overline{X})^2} = \frac{cov(x,y)}{\sigma_x^2}$$

Итого получаем коэффициенты предполагаемой модели:

$$b = \frac{cov(x,y)}{\sigma_x^2} = \frac{\sum(X_i - \overline{X}) \cdot (Y_i - \overline{Y})}{\sum(X_i - \overline{X})^2} = \frac{\overline{XY} - \overline{X} \cdot \overline{Y}}{\overline{X^2} - (\overline{X})^2} \tag{8}$$

$$a = \overline{Y} - b\overline{X} \tag{9}$$

b - наклон линии регресии (коэффициент регрессии) - абсолютный показатель силы связи.

Свойства метода МНК (результаты относительно регрессий, оцениваемых по обычному МНК):

1. $\sum e_i = 0$
2. $\overline{e} = 0$
3. $\overline{\hat{Y}} = \overline{Y}$
4. $\sum X_i \cdot e_i = 0$
5. $\sum \hat{Y}_i \cdot e_i = 0$

Уравнение регрессии всегда дополняется обязательным показателем тесноты связи. При использовании линейной регресии в качестве такого показателя выступает *линеный коэффициент корреляции* r_{xy}. Существует разные модификации формулы линейного коэффициента корреляции:

$$r = b\frac{\sigma_x}{\sigma_y} = \frac{cov(x,y)}{\sigma_x^2} \cdot \frac{\sigma_x}{\sigma_y} = \frac{\overline{YX} - \overline{Y} \cdot \overline{X}}{\sigma_x \sigma_y}, -1 \le r \le 1 \tag{10}$$

Шкала значений коэффициента корреляции (все значения берутся по модулю):

- $r \leq 0.3$ - связь слабая
- $0.3 < r \leq 0.5$ - связь умеренная
- $0.5 < r \leq 0.7$ - связь заметная
- $0.7 < r \leq 0.9$ - связь высокая
- $0.9 < r \leq 1$ - связь весьма высокая, близкая к функциональной

Следует иметь ввиду, что величина линейного коэффициента корреляции оценивает тесноту связи рассматриваемых признаков в её линейной форме. Поэтому близость абсолютной величины линейного коэффициента корреляции к нулю *еще не означает отсутствие связи* между признаками.

Для оценки качества подбора линейной функции расчитывается квадрат линейного коэффициента корреляции r_{yx}^2, называемый **коэффициентом детерминации**. Коэффициент детерминации характеризует долю дисперсии результативного признака y, объясняемую регрессией, в общей дисперсии результативного признака.

$$r_{yx}^2 = \frac{\sigma_{y,obyasn}^2}{\sigma_{y,obch}^2} \tag{11}$$

1.4.2 Качество модели: коэффициент детерминации

Цель регрессии - объяснение поведения Y. В любой выборке Y оказываетяс низким, а в других - высоким. Разброс значений Y можно описать с помощью суммы квадратов отклонений от выборочного среднего.

$$\sum (Y - \overline{Y})^2$$

Все показатели корреляции основаны на правиле сложения дисперсий $\Rightarrow$ можно разложить **общую сумму квадратов отклонений** переменной Y от среднего значения $\overline{Y}$ на две части - **"объясненную"** сумму квадратов и **"необъясненную"**.

$$\sum (Y - \overline{Y})^2 = \sum (\hat{Y} - \overline{Y})^2 + \sum (Y - \hat{Y})^2 \tag{12}$$

Данное равенство можно переписать как:

$$SS_T = SS_R + SS_E \tag{13}$$

где:

$SS_T = \sum (Y - \overline{Y})^2$ - общая сумма квадратов отклонений, **факторная сумма**

$SS_R = \sum (\hat{Y} - \overline{Y})^2$ - **сумма квадратов отклонений, объясненная** регрессией

$SS_E = \sum(Y-\hat{Y})^2 = \sum e_i^2$ - **остаточная сумма** квадратов отклонений.

Введем **коэффициент детерминации**:

$$R^2 = r^2 = \frac{\sigma^2_{y,obyasn}}{\sigma^2_{y,obch}} = \frac{SS_R}{SS_T} = 1 - \frac{SS_E}{SS_T} = 1 - \frac{\sum(Y-\hat{Y})^2}{\sum(Y-\overline{Y})^2}$$

$$R^2 = \frac{\sum(\hat{Y}-\overline{Y})^2}{\sum(Y-\overline{Y})^2} = 1 - \frac{\sum(Y-\hat{Y})^2}{\sum(Y-\overline{Y})^2} \tag{14}$$

Коэффициент детерминации- обобщающий показатель оценки качества построенного уравнения регрессии.

1.4.3 Статистическая оценка достоверности регрессионной модели

После того как найдено уравнение линейной регресии, проводится оценка значимости как уравнения в целом, так и отдельных его параметров. Оценка значимости уравнения регрессии в целом дается с помощью F-критерия Фишера. При этом выдвигается нулевая гипотеза, что коэффициент регрессии равен нулю, т.е $b = 0$ и, следовательно, фактор x не оказывает влияния на результат Y.

Выберем нулевую гипотезу, по которой мы будем оценивать качество модели (в генеральной совокупности):

$$H_0\colon r^2 = 0$$
$$H_1\colon r^2 \neq 0$$

Если же прочие факторы не влияют на результат, то Y связан с X функционально и остаточная сумма квадратов $SS_E = \sum e_i^2 = 0$. В этом случае сумма квадратов отклонений равна объясненной сумме квадратов:

$$SS_T = SS_R$$

Поскольку не все точки поля корреляции лежат на линии регрессии, то всегда имеет место их разброс как обусловленный влиянием фактора X, т.е регрессией Y по X, так и вызванный действием прочих причин(необъясненная вариация).

Так как

$$R^2 = 1 - \frac{\sum(Y-\hat{Y})^2}{\sum(Y-\overline{Y})^2} = 1 - \frac{SS_E}{SS_T},$$

то если SS_T будет больше остаточной суммы квадратов SS_E, то уравнение регрессии статистически значимо и фактор X оказывает существенное воздействие на результат Y. Это равносильно тому, что коэффициент детерминации R^2 будет приближаться к единице.

Любая сумма квадратов отклонений связана с числом степени свободы (*df - degrees of freedom*), т.е числом свободы независимого варьирования признака. Число степеней свободы связано с числом единиц совокупности n и с числом определяемых по ней констант.

При расчете объясненной или факторной суммы квадратов $\sum(\hat{Y}-\overline{Y})^2$ используются теоретические (расчетные) значения результативного признака $\hat{Y}$, найденные по линии регрессии:

$$\hat{Y} = a + bX_i$$

Сумма квадратов отклонений, обусловленных линейной регрессией (следует из формулы линейного коэффициента корреляции):

$$SS_R = \sum(\hat{Y}_i - \overline{Y})^2 = b^2 \cdot \sum(X - \overline{X})^2 \tag{15}$$

так как по формулам (11) и (14):

$$r_{yx}^2 = \frac{\sigma_{y,obyasn}^2}{\sigma_{y,obch}^2} = \frac{SS_R}{SS_T} = \frac{\sum(\hat{Y} - \overline{Y})^2}{\sum(Y - \overline{Y})^2} = b^2 \cdot \frac{\sigma_x^2}{\sigma_y^2}$$

$$\sum(\hat{Y}-\overline{Y})^2 = b^2 \cdot \frac{\sigma_x^2}{\sigma_y^2} \cdot \sum(Y-\overline{Y})^2 = b^2 \cdot \frac{\sum(X - \overline{X})^2}{\sum(Y - \overline{Y})^2} \cdot \sum(Y-\overline{Y})^2 = b^2 \cdot \sum(X-\overline{X})^2$$

Данная сумма квадратов отклонений имеет 1 степень свободы, так как зависит только от одной константы коэффициента регрессии b, следовательно:

$$df_{SS_R} = 1$$

Число степеней свободы остаточной суммы квадратов при линейной регрессии:

$$df_{SS_E} = n - 2$$

Число степеней свободы для общей суммы квадратов определяется числом единиц, и поскольку мы используем среднюю вычисленную по данным выборки, то теряем одну степень свободы, следовательно:

$$df_{SS_T} = n - 1$$

В случае линейной регрессии получаем следующее равенство:

$$n - 1 = 1 + (n - 2)$$

В общем случае:

$$df_{SS_T} = df_{SS_R} + df_{SS_E}$$

$$n - 1 = m + (n - 1 - m)$$

$$df_{SS_T} = n - 1, df_{SS_R} = m, df_{SS_E} = n - 1 - m \tag{16}$$

где m - число параметров переменных.

Разделив каждую сумму квадратов на соответствующее ей число степеней свободы, получим средний квадрат отклонений или **дисперсию на одну степень свободы**:

$$MS_R = \frac{SS_R}{df_R} = \frac{\sum(\hat{Y} - \overline{Y})^2}{m} \tag{17}$$

$$MS_E = \frac{SS_E}{df_E} = \frac{\sum(Y - \hat{Y})^2}{n - 1 - m} \tag{18}$$

$$MS_T = \frac{SS_T}{df_T} = \frac{\sum(Y - \overline{Y})^2}{n - 1} \tag{19}$$

где MS_T - общая дисперсия, MS_E - остаточная, MS_R - факторная (объясненная).

Определение дисперсии на одну степень свободы приводит дисперсии к *сравнимому виду*. Сопоставляя факторную (объясненную) и остаточную дисперсию в расчете на одну степень свободы, получим величину **F-критерия**:

$$F = \frac{MS_R}{MS_E} = \frac{\text{Factor Variance with 1 df}}{\text{Remainder variance with 1 df}} \tag{20}$$

Значение F_{table} означает максимальную величину отношения дисперсия при случайном их расхождении для данного уровня вероятности и наличия нулевой гипотезы.

В математической статистике данное распределение называется распределение *Снедекора* для (n, m) степеней свободы.

Для проверки гипотезы о значимости уравнения регресии воспользуемся следующим алгоритмом:

1. Выберем в достоверной области критический уровень значимости α. Обычно выбирают маленький уровень значимости, так как вероятность попадания в критическую область при справедливости нулевой гипотезы H_0 должна быть маленькой ($\alpha \approx 0.05$).

2. Определяется табличное критическое значение критерия Фишера F_{table}

3. Если $F > F_{table}$, то H_0 отвергается $\Rightarrow$ гипотеза о случайности природы отвергается и делается вывод о существенности связи и значимости R^2

Если нулевая гипотеза справедлива, то факторная (объясненная) и остаточная дисперсия не отличаются друг от друга. Величиная F-критерия связана с коэффициентом детерминации r^2. Факторную сумму квадратов отклонений можно представить как:

$$SS_R = \sum(\hat{Y}_i - \overline{Y})^2 = b^2 \cdot \sum(X - \overline{X})^2 = r_{yx}^2 \cdot \sigma_y^2 \cdot n \tag{21}$$

так как:

$$SS_R = \sum(\hat{Y}_i - \overline{Y})^2 = r_{yx}^2 \cdot SS_T = r_{yx}^2 \sum(Y - \overline{Y})^2 =$$

$$r_{yx}^2 \cdot \frac{1}{n} \sum(Y - \overline{Y})^2 \cdot n = r_{yx}^2 \cdot \sigma_y^2 \cdot n$$

А остаточную сумму квадратов как:

$$SS_E = \sum(Y_i - \hat{Y}_i)^2 = (1 - r_{yx}^2) \cdot \sigma_y^2 \cdot n \tag{22}$$

так как:

$$r_{xy}^2 = 1 - \frac{SS_E}{SS_T} \Rightarrow SS_E = SS_T \cdot (1 - r_{xy}^2) = (1 - r_{yx}^2) \cdot \sigma_y^2 \cdot n$$

Тогда значение F-критерия равно:

$$F = \frac{MS_R}{MS_E} = \frac{\frac{SS_R}{df_R}}{\frac{SS_E}{df_E}} = \frac{r_{yx}^2}{(1 - r_{yx}^2)} \cdot \frac{n - 1 - m}{m} \tag{23}$$

где n - число единиц в совокупности, m - число параметров при переменных.

Результаты факторного анализа обычно представлены в таблице дисперсионного анализа

Источник вариации	df	SS	MS	F-критерий
Регрессия	1	14735	14735	278
Остаток	5	265	53	1
Итого	6	15000	x	x

Таблица 1: Таблица дисперсионного анализа для примера

$F_{table} = 6.61$, $278 > 6.61$ - регресия статистически значима, $r^2 \neq 0$

1.4.4 Оценка значимости коэффициентов регрессии

В линейной регресии обычно оценивается значимость не только уравнения в целом, но и отдельных его параметров. С этой целью по каждому из параметров строится его **стандартная ошибка** (случайная ошибка коэффициента регрессии).

Выдвигается нулевая гипотеза о равенстве коэффициентов регресии в генеральной совокупности:

$$H_0 : b = 0$$

$$H_1 : b \neq 0$$

Стандартная ошибка коэффициента регрессии определяется по формуле:

$$m_b = \sqrt{\frac{MS_E}{\sum(X - \overline{X})^2}} = \sqrt{\frac{\frac{\sum(Y-\hat{Y})^2}{n-1-m}}{\sum(X - \overline{X})^2}} \tag{24}$$

Вводится t-статистика:

$$t_b = \frac{b - 0}{m_b} = \frac{b}{m_b} \sim t(n-2) \tag{25}$$

так как два параметра, то число степеней свободы равно двум и данная статистика имеет распределение Стьюдента с $n - 2$ степенями свободы.

Для проверки гипотезы о значимости коэффициента регресии воспользуемся следующим алгоритмом:

1. Выберем в достоверной области критический уровень значимости α. Обычно выбирают маленький уровень значимости, так как вероятность попадания в критическую область при справедливости нулевой гипотезы H_0 должна быть маленькой.

2. Определяется табличное критическое значение критерия Стьюдентая $t_{table}(n-2)$

3. Если $|t_b| > t_{table}$, то H_0 отвергается $\rightarrow$ гипотеза о незначимости коэффициента регрессии отвергается (параметр b не случайно отличается от нуля, и сформировался под влиянием систематически действующего фактора)

Критерий опровержения гипотезы:

$$|t_b| = \frac{b}{m_b} = \frac{b}{\sqrt{\frac{MS_E}{\sum(X-\overline{X})^2}}} > t_{table} \Leftrightarrow H_o \text{ discards} \tag{26}$$

Величина m_b называется случайной ошибкой коэффициентов регресии. Если $t_b > 3$, то параметры всегда значимы.

1.4.5 Связь F и t-критериев

F-критерий Снедекора и t-критерия Стюдента для коэффициентов регрессии взаимосвязаны. Покажем эту связь:

$$t_b^2 = \frac{b^2}{m_b^2} = \frac{b^2}{\frac{\frac{\sum(Y-\hat{Y})^2}{n-1-m}}{\sum(X-\overline{X})^2}} = \frac{b^2 \cdot \sum(X - \overline{X})^2}{\frac{\sum(Y-\hat{Y})^2}{n-1-m}} \overset{1.4.3}{=} \frac{SS_R}{\frac{\sum(Y-\hat{Y})^2}{n-1-m}} = \frac{MS_R}{MS_E} = F$$

Следовательно:

$$t_b = \sqrt{F} \tag{27}$$

1.4.6 Гипотеза о коэффициенте коррелляции

Значимость линейного коэффициента корреляции проверяется на снове величины **ошибки коэффициента корреляции** m_r:

$$m_r = \sqrt{\frac{1 - r_{yx}^2}{n - 1 - m}} \tag{28}$$

Фактическое значение t-критерия Стьюдента определяется как:

$$t_r = \frac{r}{\sqrt{1 - r_{yx}^2}} \cdot \sqrt{n - 1 - m} \tag{29}$$

$$F = \frac{r_{yx}^2}{(1 - r_{yx}^2)} \cdot (n - 1 - m)$$

Для парной регрессии:

$$t_r = \frac{r}{\sqrt{1 - r_{yx}^2}} \cdot \sqrt{n - 2} \tag{30}$$

$$F = \frac{r_{yx}^2}{(1 - r_{yx}^2)} \cdot (n - 2)$$

Следовательно F и t связаны для коэффициентов корреляции:

$$t_r = \frac{r_{xy}}{m_r}$$

$$t_r^2 = F, t_b^2 = F \Rightarrow t_r^2 = t_b^2$$

$$t_r = t_b \tag{31}$$

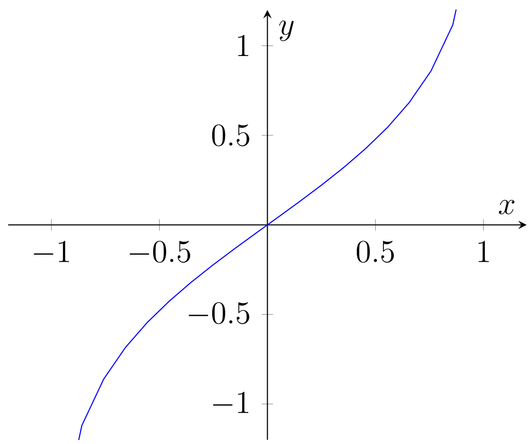
Таким образом, проверка гипотез о значимости коэффициентов регрессии и корреляции равносильна проверке гипотезы о существенности линейного уравнения регрессии. В гипотезе о корреляции: если гипотеза неверна, то зависимость является достоверной и коэффициент корреляции существенно отличен от нуля.

Рассмотренная формула оценки коэффициента корреляции работает при большом числе наблюдений и если r не близко к ± 1. Если же величина коэффициента корреляции близка к 1, то распределение его оценок отличается от нормального или распределения Стьюдента, так как величина коэффициента корреляции ограничена $[-1, 1]$.

Чтобы обойти это затруднение было предложено для оценки существенности r ввести вспомогательную величину z, связанную с коэффициентом корреляции следующим отношением:

$$z = \frac{1}{2} \cdot ln\frac{1 + r}{1 - r} \tag{32}$$

z-grid



Величина z изменяется от $-\infty$ до $+\infty$, что соответствуе пределам нормального распределения.

Стандартная ошибка величины z вычисляется по формуле:

$$m_z = \frac{1}{\sqrt{n-3}} \tag{33}$$

Далее выдвигается нулевая гипотеза H_0, которая состоит в том, что корреляция отсутствует, т.е теоретическое значение коэффициента корреляции равно 0:

$$H_0 : r_{xy} = 0, H_1 : r_{xy} \neq 0$$

Критерий опровержения гипотезы:

$$t_z = \frac{z}{m_z} = z \cdot \sqrt{n-3} \sim t(n-2) \tag{34}$$

$$t_z > t_\alpha \Leftrightarrow H_0 \text{ discards}$$

Вывод: таким образом, если H_0 отвергается, то коэффициент корреляции значимо отличен от нуля.

1.4.7 Доверительные интервалы для коэффициентов регрессии

Если коэффициенты регрессии оказываются статистически значимыми, то можно построить **доверительный интервал** для коэффициентов регрессии:

$$\delta_b = \pm t_{table} \cdot m_b$$

$$b - t_{1-\frac{\alpha}{2}} \cdot m_b \leq b \leq b + t_{1-\frac{\alpha}{2}} \cdot m_b \tag{35}$$

Также стандартную среднюю ошибку для коэффициента a можно выразить через m_b:

$$m_a = \sqrt{\frac{\sum(Y-\hat{Y})^2}{n-1-m} \cdot \frac{\sum X^2}{n \cdot (X-\overline{X})}} = m_b \cdot \sqrt{\frac{\sum X^2}{n}} \tag{36}$$

1.4.8 Использование модели парной регрессии для прогнозирования

В прогнозных расчетах по уравнению регрессии определяется предсказываемое (y_p) значение как точечный прогноз $\hat{y_x}$ при $x_p = x_k$, т.е путем подстановки в уравнение регрессии $\hat{y_x} = a + b \cdot x$ соответствующего значения x. Однако точечный прогноз явно нереален, поэтому он дополняется расчетом стандартной ошибки $\hat{y_i}$, т.е $m_{\hat{y}}$ и соответственно интервальной оценкой прогнозированного значения y^*.

Выражение для **стандартной ошибки предсказываемого по линии регрессии значения** $\hat{y}$:

$$m_{\hat{y_x}} = \sqrt{MS_E}\sqrt{\frac{1}{n} + \frac{(x_k - \overline{X})^2}{\sum(X - \overline{X})^2}} \tag{37}$$

где $\sqrt{MS_E}$ - стандартная ошибка линейной регресии. Данная формула стандартной ошибки предсказываемого значения y при заданном значении x_k и характеризует ошибку положения линии регрессии.

Величина стандартной ошибки достигает минимума при $x_k = \overline{X}$.

Для прогнозируемого значения $\hat{y}$ доверительный интервал выглядит следующим бразом:

$$\hat{y_{x_k}} \pm t_{1-\frac{\alpha}{2}} \cdot m_{\hat{y_x}}$$

$$\hat{y_{x_k}} - t_{1-\frac{\alpha}{2}} \cdot m_{\hat{y_x}} \leq \hat{y_{x_k}} \leq \hat{y_p} + t_{1-\frac{\alpha}{2}} \cdot m_{\hat{y_x}} \tag{38}$$

где:

$$\hat{y_{x_k}} = a + b \cdot x_k$$

Средняя ошибка прогнозируемого индивидуального значения составит:

$$m_y = \sqrt{MS_E}\sqrt{1 + \frac{1}{n} + \frac{(x_k - \overline{X})^2}{\sum(X - \overline{X})^2}} \tag{39}$$

Доверительный интервал для y_p - предсказываемого значения регрессии:

$$\hat{y_p} - t_\alpha m_y \leq y_p \leq \hat{y_p} + t_\alpha m_y \tag{40}$$

1.4.9 Пример использования полученных знаний

Рассмотрим выборку $\{X, Y\}$, где:

$$X = \{1, 2, 4, 3, 5, 3, 4\}$$

$$Y = \{30, 70, 150, 100, 170, 100, 150\}$$

Последовательно проведем анализ согласно изучению материала:

1. Найдем оценку параметров модели методом МНК:

Согласно формуле (7) получаем следующую систему уравнений:

$$\begin{cases} 770 = 7a + 22b \\ 2820 = 22a + 80b \end{cases}$$

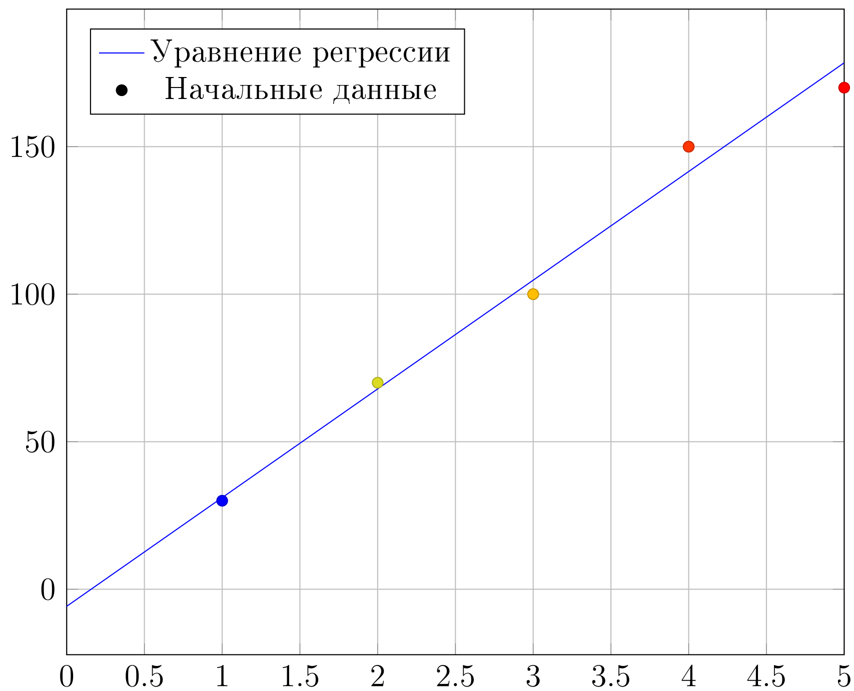
Из данной системы уравнений находим значения параметров регрессии a и b:

$$a = -5.78947; b = 36.8421$$

Можно убедиться, что все альтернативные формулы (8) дают те же значения коэффициентов линейной регрессии.

Построим график прямой

$$\hat{Y} = -5.78947 + 36.8421X$$



Линейный коэффициент корреляции по формуле (10):

$$r = 0.991189$$

Вывод: связь очень высокая и близкая к функциональной.

2. Качество модели

По формуле (13) найдем общую сумму квадратов отклонений и объясненную и необъясненную дисперсию:

$$SS_T = 15000, SS_R = 14736.8, SS_E = 263.158$$

$$SS_T = SS_R + SS_E : \text{ True}$$

По формуле (11) и (14) найдем *коэффициент детерминации*:

$$R^2 = \frac{SS_R}{SS_T} = 0.982456$$

Вывод: уравнением регрессии объясняется около 98% дисперсии результативного признака, а на долю других факторов уходит лишь 2% ее дисперсии. Чем больше коэффициент детерминации, тем меньше роль прочих факторов и, следовательно, линейная модель хорошо аппроксимирует данные и ею можно пользоваться для прогноза значений Y.

3. Проверка гипотезы о достоверности регрессионной модели

Допустим, что $H_0 : r^2 = 0$ (как следствие, коэффициент $b = 0$).

Выберем уровень значимости: $\alpha = 0.05$

Согласно формулам (17-19), высчитаем дисперсию на одну степень свободы :

$$MS_R = \frac{SS_R}{1} = 14736.8$$

$$MS_E = \frac{SS_E}{n-1-m} = \frac{263.158}{7-2} = 52.6316$$

Посчитаем статистику F-критерия:

$$F_{stat} = \frac{MS_R}{MS_E} = \frac{14736.8}{52.6316} = 280$$

Найдем табличное значение распределения Фишера-Снедекора при заданном уровне значимости:

$$F_{1-\alpha}(m, n) = F_{0.95}(1, 5) = 6.61$$

Вывод: так как $F_{stat} > F_{0.95}$, то нулевая гипотеза H_0 отвергается и делается вывод о том, что регрессия статистически значима и связь существенна.

Также можно проверить, что значение F-критерия одинаково и при других альтернативе формулы (23)(можно проверить). Дисперсионная таблица представлена на странице 10.

4. Проверка гипотезы о достоверности регрессионной модели

Проверим значимость отдельных параметров регрессии, в данном случае значимость параметра b.

Введем нулевую гипотезу о том, что параметр регрессии незначим:

$$H_0 : b = 0.$$

Выберем уровень значимости: $\alpha = 0.05$

Вычислим стандартную ошибку по формуле (24), чтобы построить t_{stat}:

$$m_b = \sqrt{\frac{52.6316}{10.8571}} = 2.20174$$

Вычислим t_{stat} по формуле (25) :

$$t_b = \frac{b}{m_b} = \frac{36.8421}{2.20174} = 16.7332$$

$$t_b = \sqrt{F} = \sqrt{280} = 16.7332$$

Вычислим табличное значение распределения Стьюдента с $(n-2)$ степенями свободы:

$$t_{table} = t_{1-\frac{\alpha}{2}} = t_{0.975} = 2.57058$$

Вывод: так как $t_b > t_{table}$, то нулевая гипотеза H_0 отвергается и делается вывод, что коэффициент линейной регрессии b статистически значим

Доверительный интервал для коэффициента b выглядит следующим образом, согласно формуле (35):

$$36.8421 - 2.57058 \cdot 2.20174 \leq b \leq 36.8421 + 2.57058 \cdot 2.20174$$

$$31.1824 \leq b \leq 42.5019$$

5. Использование модели парной регрессии для прогнозирования

Вычислим стандартную ошибку предсказываемого по линии регрессии значения $\hat{Y}$ по формуле (37):

$$m_{\hat{y_x}} = \sqrt{52.6316}\sqrt{\frac{1}{7} + \frac{(x_k - 3.14286)^2}{10.8571}}$$

Подставляя различные значения из выборки X мы можем узнать ошибку предсказываемого значения. Минимальная ошибка будет при подстановке $x_k = \overline{X} = 3.14286$:

$$m_{y_{\overline{X}}} = \sqrt{52.6316}\sqrt{\frac{1}{7}} = 2.74204$$

Построим доверительный интервал для $\hat{Y}$ при каком-то произвольном значении x_k, например $x_k = 4$. Воспользуемся формулой (38).

Сначала вычислим значение линейной регресии в точке $x_k = 4$:

$$\hat{y_4} = -5.78947 + 36.8421 \cdot 4 = 141.579$$

Затем вычислим стандартную ошибку в точке $x_k = 4$:

$$m_{\hat{y}_4} = \sqrt{52.6316}\sqrt{\frac{1}{7} + \frac{(4 - 3.14286)^2}{10.8571}} = 3.32871$$

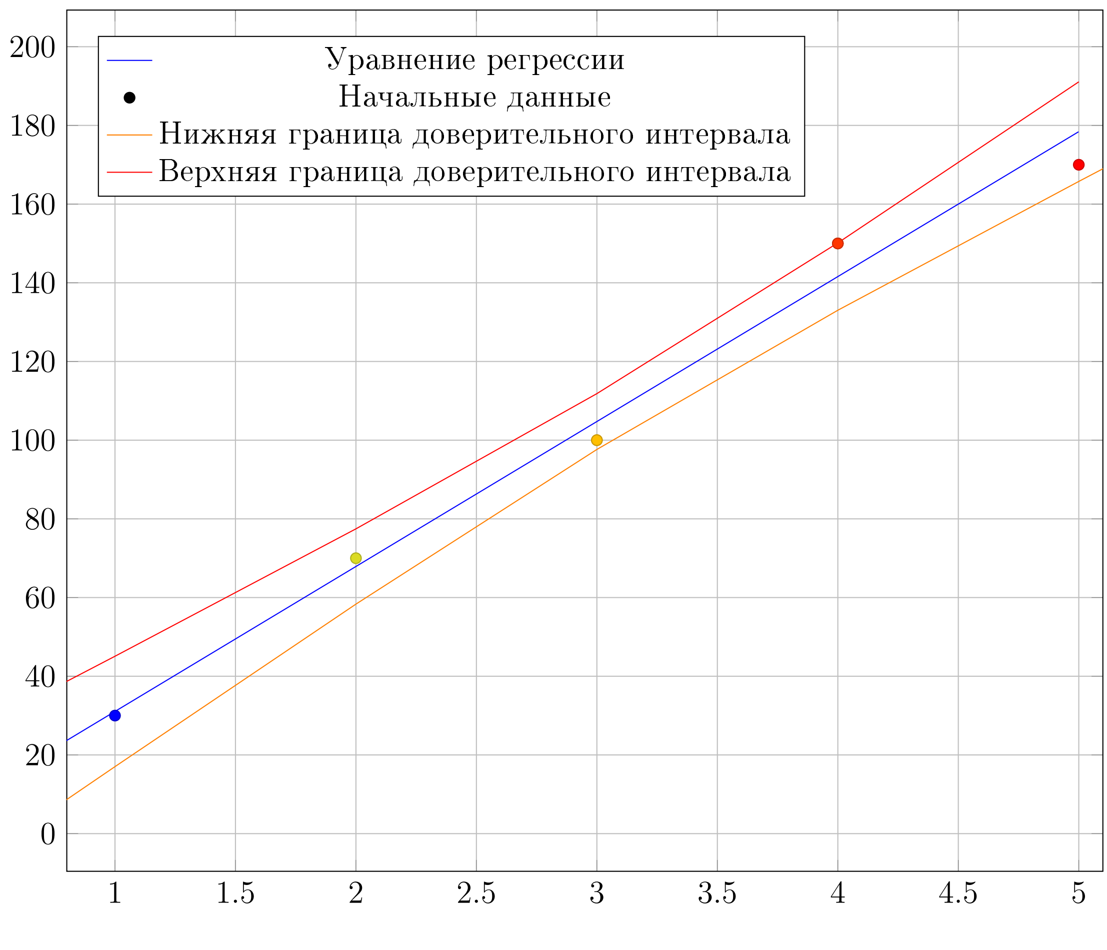
Теперь можно и построить доверительный интервал для уровня значимости $\alpha = 0.05$:

$$\hat{y}_4 - t_{0.975} \cdot m_{\hat{y}_4} \leq \hat{y}_4 \leq \hat{y}_4 + t_{0.975} \cdot m_{\hat{y}_4}$$

$$133.022 \leq \hat{y}_4 \leq 150.136$$

Значения будут удаляться от линии регрессии по гиперболе, с минимум ошибки в точке $x_k = \bar{X}$. Изобразим это на графике:

График стандартной ошибки и доверительные интервалы



Вычислим среднюю ошибку прогноза по формуле (39) и построим доверительный интвервал. Все действия аналогичны разобранному пункту.

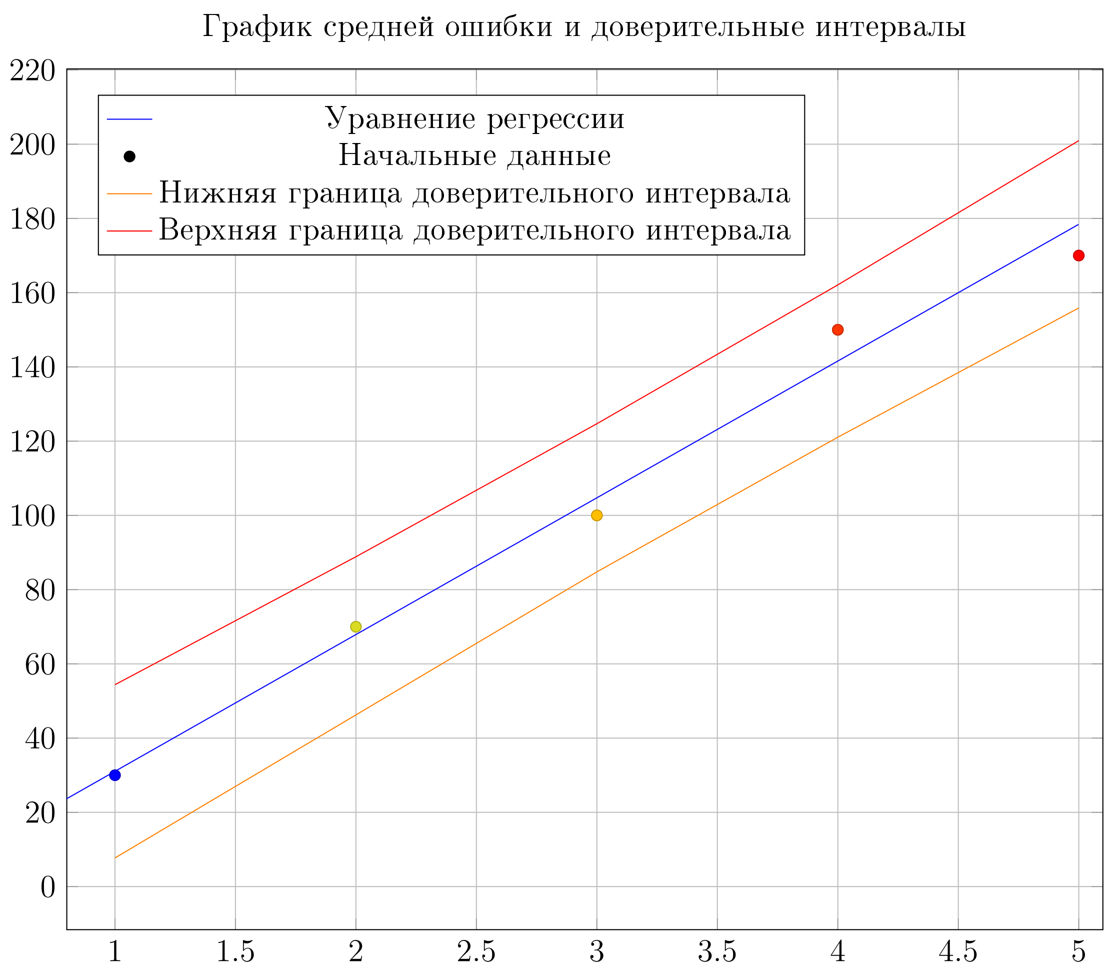
Средняя ошибка прогноза:

$$m_{\hat{y_x}} = \sqrt{52.6316}\sqrt{1 + \frac{1}{7} + \frac{(x_k - 3.14286)^2}{10.8571}}$$

Доверительный интервал для $x_k = 4$ и уровня значимости $\alpha = 0.05$ по формуле (40):

$$121.061 \leq y_{\hat{x_{k=4}}} \leq 162.097$$

Построим график средней ошибки в зависимости от наблюдений.



1.5 Нелинейная регрессия

Будут рассмотрены два вида **нелинейной регресии**:

1. Нелинейная по независимым переменным

2. Нелинейная по оцениваемым параметрам

Для 1 проводится линеаризация:

$$y = a + \frac{b}{x} = a + b \cdot \frac{1}{x} = a + b \cdot b(z) \Rightarrow y(z) - \text{linear regression}$$

При построении модели регрессии часто расчитывается показатель эластичности. Только у степенной функции показатель эластичности равен константе. Во всех остальных случаях показатель эластичности зависит от x:

$$\varepsilon = \frac{\mathrm{dy}}{\mathrm{d}x} \cdot \frac{x}{y}$$