Эконометрика

ПМ-1701

Преподаватель:

Курышева Светлана Владимировна

Санкт-Петербург
2020 г., 6 семестр

Список литературы

[1] Эконометрика: Учебник/И.И.Елисеева и др.-М.:Проспект, 2009

[2] Практикум по эконометрике: Учебное пособие/И.И.Елисеева и др.,М.:Финансы и статистика,2006

[3] Эконометрика: Учебник/В. С.Мхитарян и др.-М.:2008

[4] Доугерти К. Введение в эконометрику: Учебник. 2-е изд. / Пер. с англ. – М.: ИНФРА – М, 2007

[5] Берндт Э. Практика эконометрики: классика и современность. М.,2005

Содержание

1 Парная регрессия

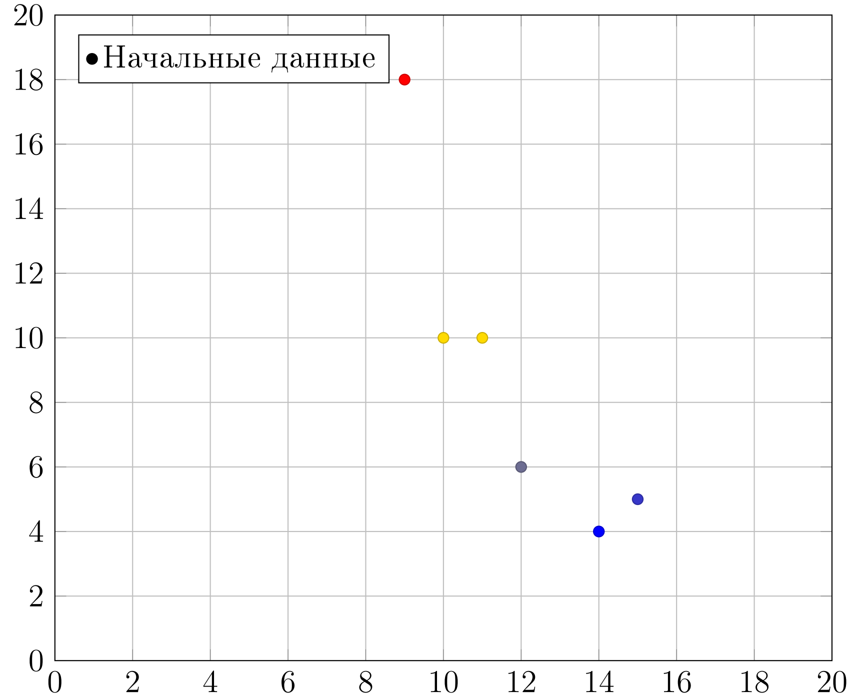
1.1 21.02.2020

Дана зависимость спроса от цены:

$$X = (15, 14, 12, 11, 9, 10)$$

$$Y = (5, 4, 6, 10, 18, 10)$$

1. Необходимо построить поле корреляции и выбрать математическую функцию.



По данной информации лучшей аппроксимации является нелинейная регрессия - степенная функция.

2. Найти линейное уравнение, используя МНК.

$$y = a + bx$$

Согласно формуле (7) получаем следующую систему уравнений:

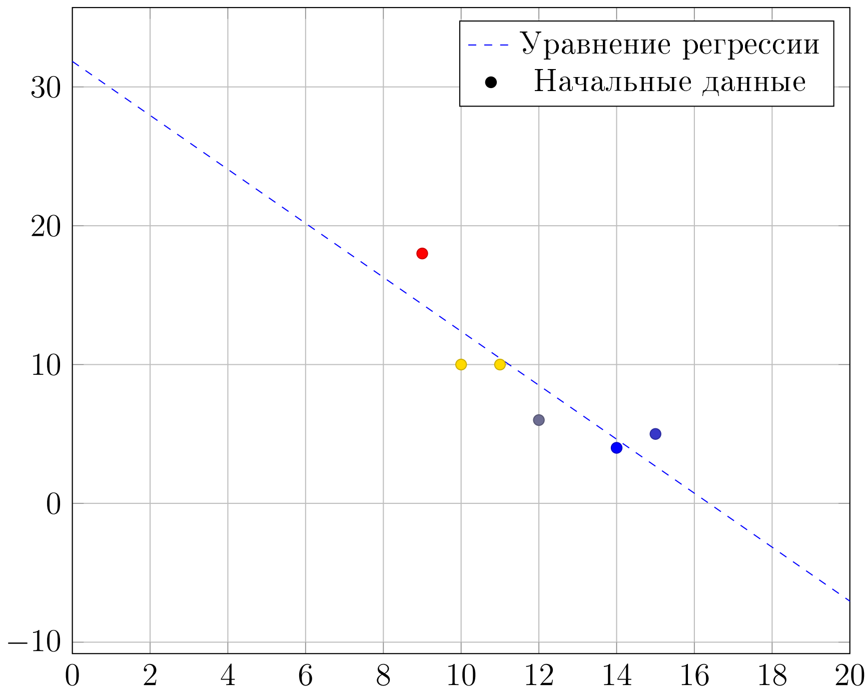
$$\begin{cases} 53 = 6a + 71b \\ 575 = 71a + 867b \end{cases}$$

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

$$a = 31.8385; b = -1.9441$$

Построим график прямой

$$\widehat{Y} = 31.8385 - 1.9441X$$



Линейный коэффициент корреляции по формуле (10):

$$r = -0.87378$$

3. Построить таблицу дисперсионного анализа:

Источник вариации	df	SS	MS	F-критерий
Регрессия (r)	1	101.417	101.417	12.9127
Остаток (e)	4	31.4161	7.85404	1
Итого (t)	5	132.833	26.5667	x

Таблица 1: Таблица дисперсионного анализа для примера

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

$$F_{1-\alpha}(m, n-1-m) = F_{0.95}(1, 4) = 7.71$$

4. Найти линейное уравнение регрессии, используя программу Excel.

1.1.1 Hometask на 28.02.2020

5. Дать интервальный прогноз спроса, для $x_p = 9$

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

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

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

$$\widehat{y_{x_k}} \pm t_{1-\frac{\alpha}{2}} \cdot m_{\widehat{y_x}}$$
$$\widehat{y_{x_k}} - t_{1-\frac{\alpha}{2}} \cdot m_{\widehat{y_x}} \le \widehat{y_{x_k}} \le \widehat{y_p} + t_{1-\frac{\alpha}{2}} \cdot m_{\widehat{y_x}} \tag{38}$$

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:

$$\widehat{y_p} - t_\alpha m_y \le y_p \le \widehat{y_p} + t_\alpha m_y \tag{40}$$

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

$$m_{\widehat{y_x}} = \sqrt{7.85404}\sqrt{\frac{1}{6} + \frac{(x_k - 11.8333)^2}{26.8333}}$$

Подставляя различные значения из выборки X мы можем узнать ошибку предсказываемого значения. Минимальная ошибка будет при подстановке $x_k = \overline{X} = 11.8333$:

$$m_{y_{\overline{X}}} = \sqrt{11.8333}\sqrt{\frac{1}{6}} = 1.14412$$

Построим доверительный интервал для $\widehat{Y}$ при каком-то произвольном значении x_k, например $x_k = 9$. Воспользуемся формулой (38).

Сначала вычислим значение линейной регресии в точке $x_k = 9$:

$$\widehat{y_9} = 31.8385 - 1.9441 \cdot 9 = 14.3416$$

Затем вычислим стандартную ошибку в точке $x_k = 9$:

$$m_{\widehat{y_9}} = \sqrt{7.85404}\sqrt{\frac{1}{6} + \frac{(9 - 11.8333)^2}{26.8333}} = 1.91278$$

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

$$\widehat{y_9} - t_{0.975} \cdot m_{\widehat{y_9}} \le \widehat{y_9} \le \widehat{y_9} + t_{0.975} \cdot m_{\widehat{y_9}}$$

$$9.0309 \le \widehat{y_9} \le 19.6523$$

Средняя ошибка прогнозируемого индивидуального значения:

$$m_y = \sqrt{MS_E}\sqrt{1 + \frac{1}{n} + \frac{(x_k - \overline{X})^2}{\sum(X - \overline{X})^2}}$$
$$= \sqrt{7.85404}\sqrt{1 + \tfrac{1}{6} + \tfrac{(9-11.8333)^2}{26.8333}} = 3.39304$$

Доверительный интервал для y_p - предсказываемого значения регрессии:

$$4.92101 \leq \widehat{y_9} \leq 23.7622$$

6. Используя Excel найти уравнение регрессии по степенной функции.

1. Степенная функция

Модель:

$$y = a \cdot x^b \cdot \varepsilon$$

Логарифмируем обе части равенства (линеаризация):

$$\ln y = \ln a + b \ln x + \ln \varepsilon$$

Замена переменных:

$$\ln y = z, \alpha_1 = \ln a, t = \ln x, \varepsilon_1 = \ln \varepsilon$$

Линейный вид:

$$z = \alpha_1 + b \cdot t + \varepsilon_1$$

В нашем случае нам нужно прологарифмировать наши ряды x и y, найти коэффициенты a и b линейной функции и перейти обратно к степенной, сделав замену $a = \mathrm{e}^{\alpha_1}$

$$X = (15, 14, 12, 11, 9, 10)$$

$$x_{new} = \ln X = (2.70805, 2.63906, 2.48491, 2.3979, 2.19722, 2.30259)$$

$$Y = (5, 4, 6, 10, 18, 10)$$

$$y_{new} = \ln Y = (1.60944, 1.38629, 1.79176, 2.30259, 2.89037, 2.30259)$$

Находим коэффициенты методом МНК:

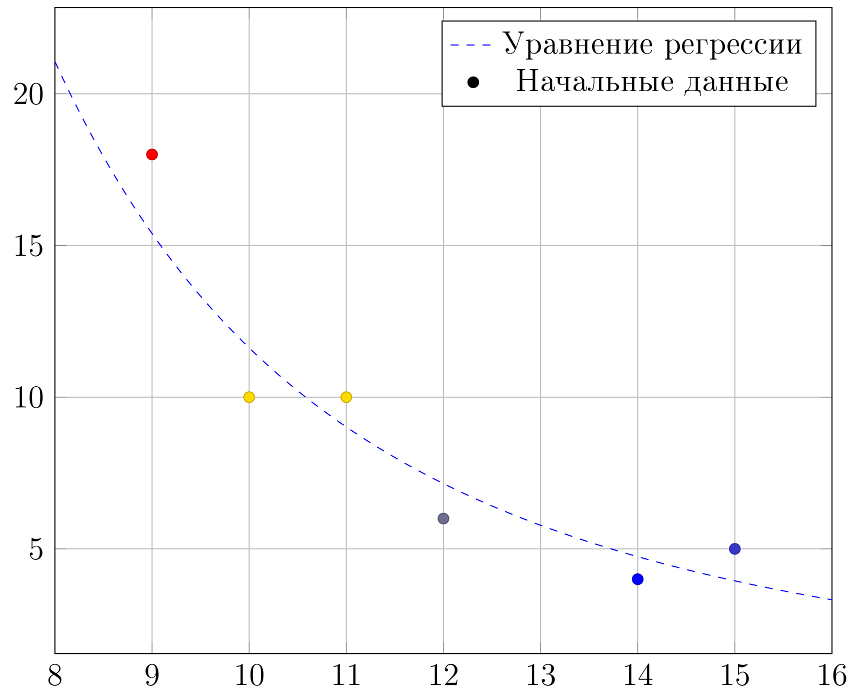
$$\alpha_1 = 8.58854, b = -2.66456$$

Исходная степенная модель имеет вид:

$$y = \mathrm{e}^{\alpha_1} x^b = 5369.78 \cdot x^{-2.66456}$$

Построим график степенной функции

$$y = 5369.78 \cdot x^{-2.66456}$$



1.2 28.02.2020

Задача 1.

Изучается зависимость спроса от цены: $Y(X)$. Имеется выборка из 5 наблюдений и по данным наблюдениям построена модель линейной регресси: $Y = 12.48 - 1.05X$. Линейный коффициент корреляции для данной модели равен $r_{xy} = -0.97206$. Дисперсия спроса $\sigma_Y^2 = 2.3336$, а дисперсия цены равна $\sigma_X^2 = 2$. Также известно, что $\bar{X} = 4, \bar{Y} = 8.28$

Найти уравнение регрессии: $X(Y)$.

Решение:

Необходимо найти уравнение $X = A + BY$.

Известна формула $|r_{yx}| = \sqrt{b \cdot B}$, поэтому:

$$B = \frac{r_{yx}^2}{b} = -0.8999$$

$$A = \overline{X} - B\overline{Y} = 4 + 0.8999 \cdot 8.28 = 11.4512$$

Соответственно, уравнение регресси принимает вид:

$$X = 11.4512 - 0.8999Y$$

Задача 2.

Коффициент регрессии $a = 4$, его стандартная ошибка $m_a = 0.8$, коффициент регрессии $b = 0.12$, его стандартная ошибка равна $m_b = 0.045$.Также даны результаты регрессии:

Источник вариации	df	SS
Регрессия (r)	1	1400
Остаток (e)	18	3600
Итого (t)	19	5000

Решение:

1. Найти общее количество наблюдений.

$$df_{SS_E} = n - 1 - m \Rightarrow n = df_{SS_E} + 1 + 1 = 20$$

2. Оценить значимость уравнения регрессии с помощью F-критерия и его параметров($\alpha = 0.05$).

$$t_b = \frac{b}{m_b} = 2.66667 > t_{0.975}(20-2) = 2.10092 - \text{значим}$$

$$t_a = \frac{a}{m_a} = 5 > t_{0.975}(20-2) = 2.10092 - \text{значим}$$

$$F_{stat} = \frac{MS_R}{MS_E} = \frac{SS_R}{m} \cdot \frac{n-1-m}{SS_E} = \frac{1400}{1} \cdot \frac{18}{3600} = 7 >$$
$$> F_{1-\alpha}(m, n-1-m) = F_{0.95}(1, 18) = 4.41 - \text{связь существенна}$$

3. Найти коффициент детерминации.

$$R^2 = \frac{SS_R}{SS_T} = \frac{1400}{5000} = 0.28$$

4. Дать интервальную оценку для коэффициента регрессии.

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

$$b - t_{1-\frac{\alpha}{2}}(n-2) \cdot m_b \leq b \leq b + t_{1-\frac{\alpha}{2}}(n-2) \cdot m_b \tag{35}$$
$$0.0254585 \leq b \leq 0.214541$$

Домашнее задание: практикум, парная регрессия, задачи 2,8 (стр.32).

1.2.1 Домашнее задание на 06.03.2020

Задание 8 (практикум)

Моделирование прибыли фирмы по уравнению $y = ab^x$ привело к следующим результатам:

$$Y = (10, 12, 15, 17, 18, 11, 13, 19)$$

$$\widehat{Y} = (11, 11, 17, 15, 20, 11, 14, 16)$$

Решение:

1. Определите ошибку аппроксимации:

$$\text{MAPE} = \frac{\sum \left| \frac{Y_i - \widehat{Y}_i}{Y_i} \right| \cdot 100\%}{n} = 9.7503$$

2. Найдите показатель тесноты связи прибыли с исследуемым в модели фактором:

$$R^2 = r^2 = \frac{\sigma^2_{y,obyasn}}{\sigma^2_{y,obch}} = \frac{SS_R}{SS_T} = 1 - \frac{SS_E}{SS_T} = 1 - \frac{\sum(Y - \widehat{Y})^2}{\sum(Y - \overline{Y})^2} = 1 - \frac{24}{79.875} = 0.699531$$

$$r_{xy} = \sqrt{R^2} = 0.836379$$

Связь является сильной.

3. Рассчитайте F-критерий Фишера:

$$F = \frac{MS_R}{MS_E} = \frac{\frac{SS_R}{df_R}}{\frac{SS_E}{df_E}} = \frac{r^2_{yx}}{(1 - r^2_{yx})} \cdot \frac{n - 1 - m}{m} = \frac{0.699531}{1 - 0.699531} \cdot \frac{8 - 1 - 1}{1} = 13.9688$$

$$F > F_{table}(m, n - 1 - m) = F_{table}(1, 6) = 5.99 \Rightarrow$$

связь между результатом и фактором существенна.

Задание 2 (практикум)

При изучении спроса на телевизоры марки N по 19 торговым точкам аналитики компании ABC выявили следующую зависимость:

$$\ln y = 10.5 - 0.8 \ln x + e$$

До проведения эксперимента администрация считала, что эластичность спроса по цене для телевизоров марки N составляет -0.9. Подтвердились ли результаты исследований? $t_b = -4.0$

Решение:

Построим доверительный интервал для коэффициента b:

$$b = -0.8 \text{т.к степенная функция}$$

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

$$t_b = \frac{b}{m_b} \Rightarrow m_b = \frac{b}{t_b} = \frac{-0.8}{-4.0} = 0.2$$
$$t_{1-\frac{\alpha}{2}}(n-2) = t_{1-\frac{\alpha}{2}}(19-2) = 2.10982$$
$$-1.22196 \leq b \leq -0.378037$$

Так как b для степенной функции является коэффициентом эластичности, то значение -0.9 поподает в доверительный интервал. Гипотеза об эластичности не отвергается.

2 Множественная регрессия

2.1 06.03.2020

Задача 1

Дана зависимость прибыли y от x_1 - единиц продукции и x_2 - производительности фондов:

$$y = (11, 17, 15, 12, 13, 8, 10, 14, 16, 18)$$
$$x_1 = (9, 15, 13, 10, 11, 7, 8, 13, 15, 20)$$
$$x_2 = (12, 20, 21, 13, 10, 6, 7, 19, 22, 18)$$

Решение:

1. Найти уравнение регрессии:

$$\begin{cases} \sum y = n \cdot a + b_1 \sum x_1 + b_2 \sum x_2 \\ \sum y \cdot x_1 = a \sum x_1 + b_1 \sum x_1^2 + b_2 \sum x_1 x_2 \\ \sum y \cdot x_2 = a \sum x_2 + b_1 \sum x_1 x_2 + b_2 \cdot \sum x_2^2 \end{cases}$$

Подставляя данные значения и решая систему, получим следующий результат:

$$\widehat{y} = 3.86985 + 0.57921 x_1 + 0.170386 x_2$$

$$\widehat{y} = 11.1274, 15.9657, 14.9777, 11.877, 11.945, 8.94663, 9.69623, 14.6369, 16.3065, 18.521$$

2. Найти R^2 - коэффициент детерминации:

$$SS_E = 3.97401, SS_R = 88.426, SS_T = 92.4$$
$$R^2 = 1 - \frac{SS_E}{SS_R} = 0.956991$$

3. Построим коррелляционную матрицу:

$$r_{y,x_1} = 0.958612, r_{y,x_2} = 0.874451, r_{x_1,x_2} = 0.786536$$

$\times$	y	x_1	x_2
y	1	0.958612	0.874451
x_1	0.958612	1	0.786536
x_2	0.874451	0.786536	1

4. Высчитывать коэффициент детерминации по корреляционной матрице:

Коэффициент детерминации может быть определен через матрицу парных коэффициентов корреляции:

$$R^2 = 1 - \frac{|\Delta r|}{\Delta r_{11}} = 0.956991$$

$$\Delta r_{11} = \begin{vmatrix} 1 & 0.786536 \\ 0.786536 & 1 \end{vmatrix}$$

5. Нормированный коэффициент множественной корреляции:

$$\hat{R}^2 = 1 - (1 - R^2)\frac{n-1}{n-m-1} = 0.944703$$

6. Множественный коэффициент детерминации:

$$r_{xy} = \sqrt{R^2} = 0.978259$$

7. Показатель частной корреляции:

При двухфакторной модели:

$$\hat{y} = a + b_1x_1 + b_2x_2$$

Частная корреляция равна:

$$r_{yx_1x_2} = \sqrt{1 - \frac{1 - R^2_{yx_1x_2}}{1 - r^2_{yx_2}}} = 0.904016$$

$$r_{yx_2x_1} = \sqrt{1 - \frac{1 - R^2_{yx_1x_2}}{1 - r^2_{yx_1}}} = 0.68516$$

$$R^2_{yx_1x_2} = R^2$$

8.Оценка значимости модели множественной регрессии:

$$F_{x_1} = \frac{R^2_{yx_1x_2} - r^2_{yx_2}}{1 - R^2_{yx_1x_2}} \cdot (n-3) = 31.3026$$

$$F_{x_2} = \frac{R^2_{yx_1x_2} - r^2_{yx_1}}{1 - R^2_{yx_1x_2}} \cdot (n-3) = 6.19373$$

Оба значения больше $F(m, n-m-1) = F(1,7) = 5.59$, поэтому включение в модель фактора x_1 после фактора x_2 статистически оправдано.

9. Дать интервальный прогноз предполагая, что $x_1 = 16, x_2 = 21$

9.1 Определим точечный прогноз для данных:

Стандартная ошибка регрессии:

$$S = \sqrt{\frac{\sum \xi_i^2}{n-m-1}} = 0.753469$$

Ошибка прогнозного значения функции регрессии получим по формуле:

$$m_{\hat{y}_p} = \sqrt{\frac{\sum \xi_i^2}{n-m-1}(X_p^T(X^TX)^{-1}X_p)} = S\sqrt{X_p^T(X^TX)^{-1}X_p}$$

где:

$$X_p = \begin{pmatrix} 1 \\ 16 \\ 21 \end{pmatrix}$$

$$X_p^T = \begin{pmatrix} 1 & 16 & 21 \end{pmatrix}$$

Матрица X^TX составляется из коэффициентов в правой части при поиске уравнения регрессии:

$$\begin{cases} \sum y = n \cdot a + b_1 \sum x_1 + b_2 \sum x_2 \\ \sum y \cdot x_1 = a \sum x_1 + b_1 \sum x_1^2 + b_2 \sum x_1 x_2 \\ \sum y \cdot x_2 = a \sum x_2 + b_1 \sum x_1 x_2 + b_2 \cdot \sum x_2^2 \end{cases}$$

$$\begin{Bmatrix} n & \sum x_1 & \sum x_2 \\ \sum x_1 & \sum x_1^2 & \sum x_1 x_2 \\ \sum x_2 & \sum x_1 x_2 & \sum x_2^2 \end{Bmatrix}$$

Высчитывая ошибка прогнозного значения функции регрессии, получим:

$$m_{\hat{y}_p} = 0.361065$$

Предсказанное значение:

$$\hat{y}(16, 21) = 16.7153$$

97.5% квантиль распределения Стьюдента с $n-m-1 = 10-2-1 = 7$ степенями свободы равен $t_{97.5\%}(7) = 2.36462$

Предельная ошибка:

$$\delta_{\hat{y}_p} = t_{table} m_{\hat{y}_p} = 0.853782$$

Интервальная оценка прогнозного значения функции регрессии определяется по формуле:

$$\hat{y} - \delta_{\hat{y}_p} \leq \hat{y_{gen}} \leq \hat{y} + \delta_{\hat{y}_p}$$

$$15.8615 \leq \hat{y_{gen}} \leq 17.5691$$

9.2 Доверительный интервал для индивидуального прогнозного значения:

$$\hat{y} - \delta_{\hat{y}_p} \leq \hat{y_{pred}} \leq \hat{y} + \delta_{\hat{y}_p}$$

$$m_{y_i} = \sqrt{\frac{\sum \xi_i^2}{n-m-1}(1 + X_p^T (X^T X)^{-1} X_p)} = S\sqrt{1 + X_p^T (X^T X)^{-1} X_p}$$

Путем подстановки значений, получаем:

$$14.7396 \leq \hat{y}_{pred} \leq 18.691$$

Домашнее задание: построить по этим же данным степенную модель и дать прогноз для тех же значений.

2.2 Домашнее задание на 13.03.2020

Дана зависимость прибыли y от x_1 - единиц продукции и x_2 - производительности фондов:

$$y = (11, 17, 15, 12, 13, 8, 10, 14, 16, 18)$$

$$x_1 = (9, 15, 13, 10, 11, 7, 8, 13, 15, 20)$$

$$x_2 = (12, 20, 21, 13, 10, 6, 7, 19, 22, 18)$$

Построить степенную модель и дать прогноз для тех же значений:

Решение:

1. Степенная модель выглядит следующим образом:

Модель:

$$y = a \cdot x_1^b \cdot x_2^c \cdot \varepsilon$$

Логарифмируем обе части равенства (линеаризация):

$$\ln y = \ln a + b \ln x_1 + c \ln x_2 + \ln \varepsilon$$

Получаем уравнение, решая методом МНК:

$$\ln y = 0.750122 + 0.592206 \ln x_1 + 0.141373 \ln x_2 + \ln \varepsilon$$

В степенном виде это выглядит, как:

$$y = e^{0.750122} + x_1^{0.592206} + x_2^{0.141373}$$

В данной моделе можно сравнивать коэффициенты эластичности между собой, так как степенная функция.

2. Доверительный интервал для прогнозного значения:

Стандартная ошибка:

$$S = 0.053857$$

Прогнозное значение по модели:

$$\hat{y}(x_1 = 16, x_2 = 21) = e^{0.750122} + x_1^{0.592206} + x_2^{0.141373} = 16.8185$$

$$\log \hat{y} = 2.82248$$

Высчитываем среднюю ошибку прогноза. Берем вектор $X_p = (1, \ln 16, \ln 21)$ и матрицу для регрессии, приведенной из степенноого вида к линейному.

$$m_{y_i} = 0.0593691$$

Предельная ошибка:

$$\delta_{\hat{y}_p} = t_{table}(n - 1 - m)m_{y_i} = 2.36462 \cdot 0.0593691 = 0.140386$$

Прогноз для логарифмов:

$$\log \hat{y} - \delta_{\hat{y}_p} \leq \hat{\ln} y_{pred} \leq \log \hat{y} + \delta_{\hat{y}_p}$$

$$2.68209 \leq \hat{\ln} y_{pred} \leq 2.96286$$

Прогноз для степенной функции (переходим от логаримов:)

$$14.6157 \leq \hat{y}_{pred} \leq 19.3533$$

2.3 13.03.2020

Продолжаем работать с данными из задачи, рассмотренной на предыдущих уроках.

10. Оценить наличие или отсутствие гетероскедастичности:

Тесты и критерии гетероскедастичности:

1. Иногда график позволяет предположить отсутствие гетероскедастичности.

2. Тест ранговой корреляции Спирмэна. Остатки рассматриваются по модулю, устанавливается их зависимость от значений фактора x. Если $|e_i|$ и x_i будут коррелированы, то делается вывод о гетероскедатичности.

$$R = 1 - \frac{6 \sum d^2}{n \cdot (n^2 - 1)}$$

$$t_R < t_\alpha$$

Проведем данный тест для наших данных. Используем модель множественной линейной регрессии, найденную на занятии 06.03.2020:

$$y = (11, 17, 15, 12, 13, 8, 10, 14, 16, 18)$$
$$x_1 = (9, 15, 13, 10, 11, 7, 8, 13, 15, 20)$$
$$x_2 = (12, 20, 21, 13, 10, 6, 7, 19, 22, 18)$$

$$\widehat{y} = 3.86985 + 0.57921x_1 + 0.170386x_2$$

$\widehat{y} = 11.1274, 15.9657, 14.9777, 11.877, 11.945, 8.94663, 9.69623, 14.6369, 16.3065, 18.521$

Остатки модели:

$$y - \widehat{y} = (-0.127369, 1.03429, 0.0223198, 0.123035, 1.05498,$$

$$-0.946634, 0.30377, -0.636909, -0.306486, -0.520994)$$

Возьмем модуль остатков:

$$|y - \widehat{y}| = (0.127369, 1.03429, 0.0223198, 0.123035, 1.05498, 0.946634,$$

$$, 0.30377, 0.636909, 0.306486, 0.520994)$$

Перейдем к рангам в данном ряду и в рядах x_1, x_2:

$$r_{res} = (8, 2, 10, 9, 1, 3, 7, 4, 6, 5)$$

$$r_{x_1} = (8, 2.5, 4.5, 7, 6, 10, 9, 4.5, 2.5, 1)$$
$$r_{x_2} = (7, 3, 2, 6, 8, 10, 9, 4, 1, 5)$$

Посчитаем квадрат разности для рангов остатков и рангов временных рядов и посчитаем статистику Кэндалла для проверки гетероскедастичности x_1:

$$\rho = 1 - \frac{6\sum d^2}{n \cdot (n^2 - 1)} = 1 - \frac{6 \cdot 141}{10 \cdot 99} = 0.145455$$

Статистика t_R высчитывается следующим образом:

$$t_R = \frac{\rho}{\sqrt{\frac{1-\rho^2}{n-1-m}}} = 0.41583$$

$$t_{1-\frac{\alpha}{2}}(n - 1 - m) = t_{97.5\%}(7) = 2.306$$
$$t_R < t_{1-\frac{\alpha}{2}}(n - 1 - m)$$

Следовательно, остатки не обладают гетероскедастичностью, а обладают гомоскедастичностью. Для x_2 считается аналогично:

$$t_R = -0.650828 < t_{1-\frac{\alpha}{2}}(n-1-m)$$

Остатки x_2 также не обладают гетероскедастичностью.

3. Тест Гольдфельда-Квандта

Разберем применение данного метода на основе задачи.

Задача 2 (практикум стр.123)

По 20 наблюдениям была построена модель зависимости расходов на питание:

$$\hat{y} = 20.84 + 0.44x, r_{yx}^2 = 0.916$$

Даны значения x и остаков e_i:

Применим Тест Гольдфельда-Квандта. Для начала запишем в таблицу значения x, e_i, $\hat{y}$, полученное из уравнения регрессии и $y = \hat{y}_i + e_i$.

После этого исключим $n = 4$ центральных наблюдений и разобьем совокупность наших данных на две части:

1. Со значениями x, ниже центральных
2. Со значениям x, выше центральных

Заметим, что в исходных данных, значения x должны быть расположены по возрастанию. В каждой части осталось по 8 наблюдений. Найдем уравнение регрессии по каждой части:

$$\hat{y} = 5.795 + 0.629927 \cdot x, F_{stat_1} = 72.848$$

$$\hat{y} = 38.6056 + 0.364033 \cdot x, F_{stat_2} = 26.2807$$

Находим сумму квадратов остатков для каждой группы:

$$sum_1 = 68.8251$$

$$sum_2 = 1650.31$$

Найдем соотношение:

$$R = \frac{sum_2}{sum_1} = 23.9784$$

Сравним эту величину с табличным значением F-критерия:

$$F(n-1-m, n-1-m) = F(8-1-1, 8-1-1) = F(6,6) = 4.28$$

$$R > F(6,6)$$

Следовательно, делаем вывод о гетероскедастичности остатков.

2.4 Домашнее задание на 20.03.2020

4. Тест Парка:

Если остатки зависят от переменной, значит их дисперсия непостоянна, а значит они гетероскедастичны. Если есть гомоскедастичность - остатки ни от чего не зависят

Строится регрессия вида:

$$\ln e^2 = a + b \ln x$$

Уравнение регрессии для наших данных:

$$\ln e^2 = -0.887929 + 1.03412 \cdot \ln x$$

$$R^2 = 1 - \frac{SS_E}{SS_T} = 0.163975$$

$$F_{stat} = 3.53046 < F_{m=1,n-1-m=18} = 4.41$$

По этим данным можно сделать вывод, что наши остатки гомоскедастичны - R^2 не высок, регрессия в принципе не значима, поэтому остатки гомоскедастичны по тесту Парка.

5. Тест Глейзера:

В тесте Глейзера мы оцениваем значимость абсолютных значений остатков от значений фактора x в виде функции

$$|e| = a + b \cdot x^c$$

где в качестве c задается какое-то значение степени. По данным строится регрессия, далее высчитываются R^2 и $t - stat$ для параметров регрессии и на основании этих значений выдается вердикт о гетероскедастичности.

В нашем случае:

Для c=1:

$$\hat{y} = 4.09 + 0.04 \cdot x$$

$$R^2 = 0.29, t_{stat:1,2} = 1.71, 2.76$$

Видим, что качество модели плохое, а один из коэффициентов регрессии не значим. Попробуем от $c = 2$:

$$\hat{y} = 6.97 + 0.0001 \cdot x$$

$$R^2 = 0.247916, t_{stat:1,2} = 3.95, 2.43$$

Коэффициенты стали значимыми.

Полученная величина зависит от остатков - гетероскедастичны.