

Лабораторная работа #3

Машинное обучение. Регрессионный анализ

Цель: изучить основы методов Machine Learning в контексте задачи множественного регрессионного анализа, приобрести навыки работы с методами Machine Learning в системе STATISTICA StatSoft, осуществить обработку методами Machine Learning индивидуального набора данных и интерпретацию результатов

1 Ход работы

- 1) изучить теоретические сведения
- 2) приобрести навыки работы с методами Machine Learning в контексте задачи множественного регрессионного анализа в системе STATISTICA StatSoft, реализуя приведенный ниже пример
- 3) на основе приобретенных практических навыков осуществить все этапы обработки методами Machine Learning в контексте задачи множественного регрессионного анализа и интерпретацию результатов согласно варианту индивидуального задания
- 4) оформить отчет и подготовиться к защите лабораторной работы по полученным результатам и контрольным вопросам

2 Содержание отчета и требования к его оформлению

- 1) отчет оформляется в печатном виде
- 2) отчет содержит титульный лист, исходные данные, результаты выполнения этапов обработки данных в виде скриншотов и обязательных комментариев по ходу выполнения работы, выводы
- 3) к отчету прилагается файл исходных данных *.sta и файл проекта в электронном виде с целью осуществления выборочного контроля.

3 Варианты исходных данных

- исходные данные - в файлах Fish for vars 1-5.xls, Real-estate for vars 6-14.xls. В соответствии с вариантом задания создать собственный файл данных и далее работать с полученными данными.

4 Краткие теоретические сведения

Регрессионный анализ (regression analysis) – это метод изучения статистической взаимосвязи между одной зависимой количественной переменной от одной или нескольких независимых количественных переменных. Зависимая переменная в регрессионном анализе называется результирующей, а переменные факторы – предикторами или объясняющими переменными.

Взаимосвязь между средним значением результирующей переменной y и средними значениями предикторов X выражается в виде уравнения регрессии. Уравнение регрессии – математическая функция, которая подбирается на основе исходных статистических данных зависимой и объясняющих переменных. Чаще всего используется линейная функция. В этом случае говорят о линейном регрессионном анализе.

Регрессионный анализ очень тесно связан с корреляционным анализом. В корреляционном анализе исследуется направление и теснота связи между количественными переменными. В регрессионном анализе исследуется форма зависимости между количественными переменными. Т.е. фактически оба метода изучают одну и ту же взаимосвязь, но с разных сторон, и дополняют друг друга. На практике корреляционный анализ выполняется перед регрессионным анализом. После доказательства наличия взаимосвязи методом корреляционного анализа можно выразить форму этой связи с помощью регрессионного анализа.

Цель регрессионного анализа – с помощью уравнения регрессии предсказать ожидаемое среднее значение результирующей переменной.

Математическое определение регрессии. Строго регрессионную зависимость можно определить следующим образом. Пусть Y, X_1, X_2, \dots, X_n — случайные величины с заданным совместным распределением вероятностей. Если для каждого набора значений $X_1=x_1, X_2=x_2, \dots, X_n=x_n$ определено условное математическое ожидание

$y(x_1, x_2, \dots, x_n) = E(Y \mid X_1=x_1, X_2=x_2, \dots, X_n=x_n)$ (уравнение линейной регрессии в общем виде),

то функция $y(x_1, x_2, \dots, x_n)$ называется *регрессией величины Y по величинам X_1, X_2, \dots, X_n* , а её график — *линией регрессии Y по X_1, X_2, \dots, X_n* , или уравнением регрессии.

Зависимость Y от X_1, X_2, \dots, X_n проявляется в изменении средних значений Y при изменении X_1, X_2, \dots, X_n . Хотя при каждом фиксированном наборе значений $X_1=x_1, X_2=x_2, \dots, X_n=x_n$ величина Y остаётся случайной величиной с определённым рассеянием.

Для выяснения вопроса, насколько точно регрессионный анализ оценивает изменение Y при изменении X_1, X_2, \dots, X_n , используется *средняя величина дисперсии Y* при разных наборах значений X_1, X_2, \dots, X_n (фактически речь идет о мере рассеяния зависимой переменной вокруг линии регрессии).

В зависимости от типа связи между переменными выделяют несколько моделей уравнений регрессии:

1. Парная линейная регрессия

$$y = \alpha + \beta x + \varepsilon$$

2. Множественная линейная регрессия

$$y = \beta_0 + \beta_1 x_1 + \dots + \beta_n x_n + \varepsilon$$

3. Полиномиальное уравнение

$$y = \beta_0 + \beta_1 x + \beta_2 x^2 + \dots + \beta_n x^n + \varepsilon$$

4. Степенное уравнение

$$y = \beta_0 x_1^{\beta_1} \dots x_n^{\beta_n} + \varepsilon$$

5. Показательное уравнение

$$y = \beta_0 \beta_1^{x_1} \dots \beta_n^{x_n} + \varepsilon \quad (\beta_i > 0, \beta_i \neq 1)$$

6. Экспоненциальное уравнение

$$y = e^{\beta_0 + \beta_1 x_1 + \dots + \beta_n x_n} + \varepsilon$$

7. Логарифмическое уравнение

$$y = \beta_0 + \beta_1 \ln x_1 + \dots + \beta_n \ln x_n + \varepsilon$$

8. Гиперболическое уравнение

$$y = \beta_0 + \frac{\beta_1}{x_1} + \dots + \frac{\beta_n}{x_n} + \varepsilon$$

Здесь α , β , β_i – коэффициенты уравнений регрессии, подлежащие расчету и оценке; ε – случайная составляющая.

Метод наименьших квадратов (расчёт коэффициентов). На практике линия регрессии чаще всего ищется в виде линейной функции (см. выше линейное уравнение регрессии, множественное уравнение линейной регрессии), наилучшим образом приближающей искомую кривую. Делается это с помощью **метода наименьших квадратов**, когда минимизируется сумма квадратов отклонений реально наблюдаемых Y от их оценок (имеются в виду оценки с помощью прямой линии, претендующей на то, чтобы представлять искомую регрессионную зависимость):

$$\sum_{k=1}^M (Y_k - \hat{Y}_k)^2 \rightarrow \min$$

(M — объём выборки). Этот подход основан на том известном факте, что фигурирующая в приведённом выражении сумма принимает минимальное значение именно для того случая, когда $Y = y(x_1, x_2, \dots, x_n)$.

Для решения задачи регрессионного анализа методом наименьших квадратов вводится понятие *функции невязки*:

$$\sigma(\bar{b}) = \frac{1}{2} \sum_{k=1}^M (Y_k - \hat{Y}_k)^2$$

Условие минимума функции невязки для случая множественной линейной регрессии:

$$\left\{ \begin{array}{l} \frac{d\sigma(\bar{b})}{db_i} = 0 \\ i = 0 \dots N \end{array} \right\} \Leftrightarrow \left\{ \begin{array}{l} \sum_{i=1}^M y_i = \sum_{i=1}^M \sum_{j=1}^N b_j x_{i,j} + b_0 M \\ \sum_{i=1}^M y_i x_{i,k} = \sum_{i=1}^M \sum_{j=1}^N b_j x_{i,j} x_{i,k} + b_0 \sum_{i=1}^M x_{i,k} \\ k = 1 \dots N \end{array} \right.$$

Рассмотрим более детально последнее условие. Условием минимизации функции невязки является равенство нулю ее производной, взятой по коэффициентам уравнения регрессии (см. слева). Отсюда, решая полученные уравнения, можно получить систему для нахождения коэффициентов (см. справа).

Для построения адекватной регрессионной модели требуется решить следующие **основные задачи регрессионного анализа**:

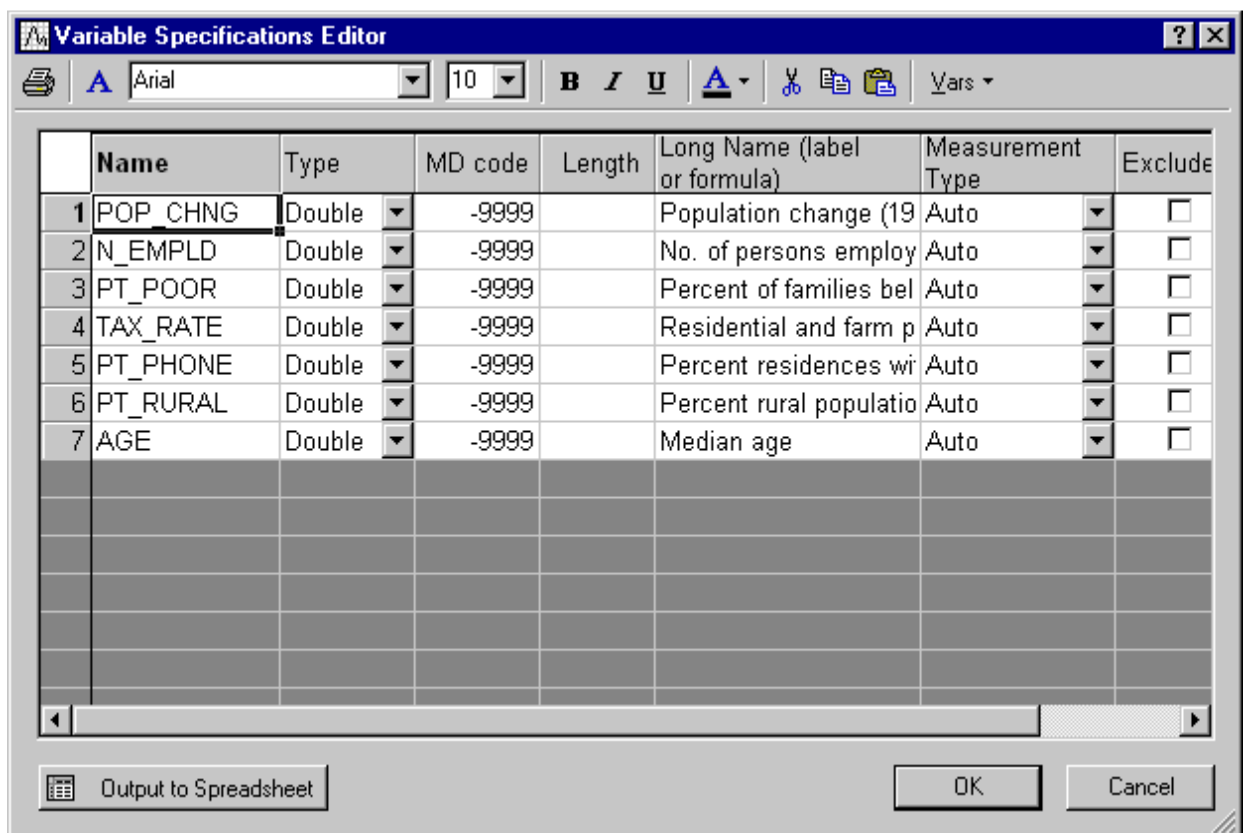
- определения вида и формы зависимости;
- оценка параметров уравнения регрессии;
- проверка значимости уравнения регрессии;
- проверка значимости отдельных коэффициентов уравнения;
- построение интервальных оценок коэффициентов;
- исследование характеристик точности модели;
- построение точечных и интервальных прогнозов результирующей переменной.

Как и корреляционный анализ, регрессионный анализ отражает только количественные зависимости между переменными. Причинно-следственные зависимости регрессионный анализ не отражает. Гипотезы о причинно-следственной связи переменных должны формулироваться и обосновываться исходя из теоретического анализа содержания изучаемого явления.

5. Пример использования регрессионного анализа *STATISTICA* в обработке данных переписи

Файл данных. Этот пример основан на файле данных Poverty.sta. Откройте этот файл данных, выбрав «Открыть примеры» в меню «Файл» (классические меню) или выбрав «Открыть примеры» в меню «Открыть» на вкладке «Главная» (полоса ленты); он находится в папке Datasets. Данные основаны на сравнении данных переписей 1960 и 1970 годов по случайному выбору 30 округов. Названия округов были введены в качестве названий случаев (строк данных).

Информация для каждой переменной указана в Редакторе спецификаций переменных Variable Specifications Editor (доступ к которому можно получить, выбрав Все спецификации переменных All Variable Specs в меню Данные).



Задача исследований - анализ коррелятов бедности, то есть переменных, которые лучше всего предсказывают процент семей за чертой бедности в округе. Таким образом, вы будете рассматривать переменную 3 (Pt_Poor) как зависимую или критериальную переменную, а все остальные переменные как независимые или предикторы.

Начало анализа. Выберите «Множественная регрессия» в меню «Статистика». Укажите уравнение регрессии, нажав кнопку «Переменные» в диалоговом окне «Множественная линейная регрессия» — вкладка «Быстрая», чтобы отобразить диалоговое окно выбора переменных. Выберите PT_POOR в качестве зависимой переменной и все остальные переменные в файле данных из списка независимых переменных, а затем нажмите кнопку ОК. Кроме того, в диалоговом окне «Множественная линейная регрессия» — вкладка «Дополнительно» установите флажок «Просмотреть описательную статистику, матрицу корреляции».



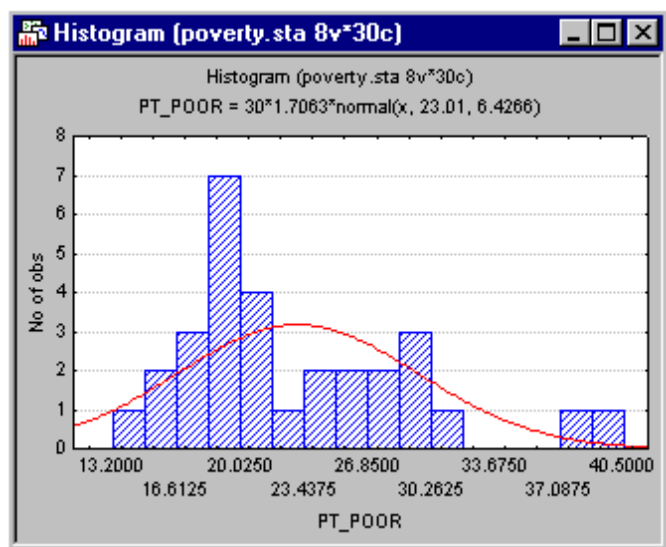
Теперь нажмите кнопку «ОК» в этом диалоговом окне, и отобразится диалоговое окно «Просмотр описательной статистики». Здесь вы можете просмотреть средние значения и стандартные отклонения, корреляции и ковариации между переменными. Обратите внимание, что это диалоговое окно также доступно практически из всех последующих диалоговых окон Множественной регрессии, поэтому вы всегда можете вернуться, чтобы просмотреть описательную статистику для конкретных переменных. Кроме того, доступно множество графиков.

Распределение переменных. Сначала изучите распределение зависимой переменной Pt_Poor по округам. Нажмите кнопку «Средние значения и стандартные отклонения», чтобы отобразить эту электронную таблицу.

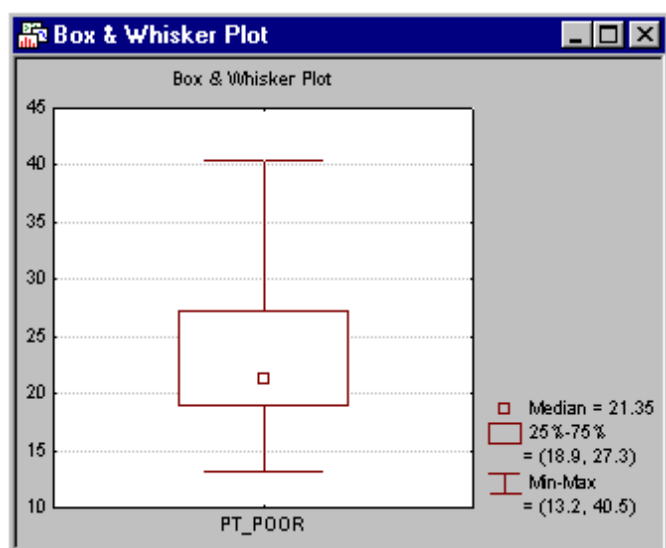
Data: Means and Standard ...			
Variable	Means and Standard De		
	Means	Std.Dev.	N
POP_CHNG	7.867	10.332	30
N_EMPLD	1548.667	2038.386	30
TAX_RATE	0.719	0.203	30
PT_PHONE	74.833	10.007	30
PT_RURAL	70.727	24.022	30
AGE	30.280	2.885	30
PT_POOR	23.010	6.427	30

Выберите «Гистограммы» в меню «Графики», чтобы создать следующую гистограмму переменной PT_POOR. В диалоговом окне «2D-гистограммы» — вкладка «Дополнительно» в группе «Интервалы» выберите кнопку выбора «Категории», введите 16 в соответствующее поле редактирования и нажмите кнопку «ОК». В диалоговом окне выбора переменных выберите PT_POOR и нажмите ОК. Как вы можете видеть на следующем изображении, распределение этой переменной несколько отличается от нормального распределения. Коэффициенты корреляции могут быть существенно

завышены или занижены, если в данных присутствуют экстремальные выбросы. Однако, даже несмотря на то, что в двух округах (два крайних правых столбца) процент семей, живущих за чертой бедности, выше, чем можно было бы ожидать в соответствии с нормальным распределением, они все же кажутся достаточно «в пределах диапазона».



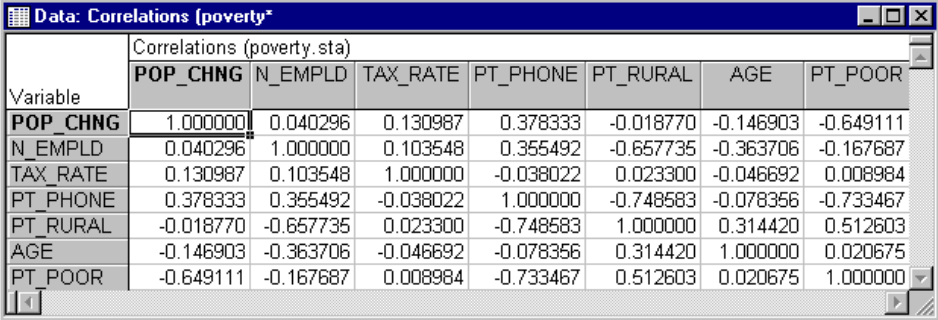
Это решение несколько субъективно; эмпирическое правило состоит в том, что нужно беспокоиться, если наблюдение (или наблюдения) выходит за пределы ± 3 -кратного среднего стандартного отклонения. В этом случае целесообразно повторить критический анализ с выбросами и без них, чтобы убедиться, что они не оказали серьезного влияния на картину взаимокорреляций. Чтобы просмотреть распределение этой переменной, нажмите кнопку График с прямоугольниками и усами в диалоговом окне «Просмотр описательной статистики» — вкладка «Дополнительно». В диалоговом окне выбора переменной выберите переменную Pt_Poor и нажмите ОК. Выберите кнопку выбора Median/Quart/Range в диалоговом окне Box-Whisker Type, а затем нажмите кнопку ОК, чтобы создать box and whisker график.



(Обратите внимание, что конкретный метод вычисления медианы и квартилей можно настроить «для всей системы» ("system-wide") на панели «Общие параметры» диалогового окна «Параметры».)

Диаграммы рассеяния. Если имеются априорные гипотезы о взаимосвязи между конкретными переменными в этот момент, может быть полезно построить соответствующую диаграмму рассеяния. Например, посмотрите на взаимосвязь между изменением численности населения и процентом семей, живущих за чертой бедности. Кажется разумным предсказать, что бедность приведет к миграции за границу; таким образом, должна существовать отрицательная корреляция между процентом населения, живущего за чертой бедности, и изменением численности населения.

Вернитесь в диалоговое окно «Просмотр описательной статистики» и нажмите кнопку «Корреляции» на вкладке «Быстрая» или «Дополнительно», чтобы отобразить электронную таблицу с матрицей корреляции.

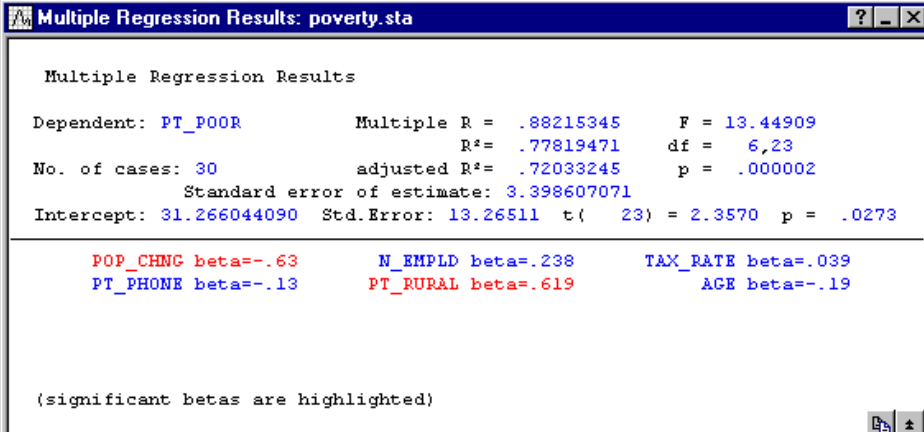


Variable	POP_CHNG	N_EMPLD	TAX_RATE	PT_PHONE	PT_RURAL	AGE	PT_POOR
POP_CHNG	1.000000	0.040296	0.130987	0.378333	-0.018770	-0.146903	-0.649111
N_EMPLD	0.040296	1.000000	0.103548	0.355492	-0.657735	-0.363706	-0.167687
TAX_RATE	0.130987	0.103548	1.000000	-0.038022	0.023300	-0.046692	0.008984
PT_PHONE	0.378333	0.355492	-0.038022	1.000000	-0.748583	-0.078356	-0.733467
PT_RURAL	-0.018770	-0.657735	0.023300	-0.748583	1.000000	0.314420	0.512603
AGE	-0.146903	-0.363706	-0.046692	-0.078356	0.314420	1.000000	0.020675
PT_POOR	-0.649111	-0.167687	0.008984	-0.733467	0.512603	0.020675	1.000000

Корреляции между переменными также можно отобразить на матричной диаграмме рассеяния. Матричную диаграмму рассеяния выбранных переменных можно создать, нажав кнопку Матричная диаграмма корреляций в диалоговом окне «Просмотр описательной статистики» — вкладка «Дополнительно», а затем выбрав нужные переменные.

Определение множественной регрессии. Теперь нажмите кнопку «ОК» в диалоговом окне «Просмотр описательной статистики», чтобы выполнить регрессионный анализ и отобразить диалоговое окно «Результаты множественной регрессии». Будет выполнена стандартная регрессия (которая включает точку пересечения).

Просмотр результатов. Поле «Сводка» в верхней части диалогового окна «Результаты множественной регрессии» отображается ниже. В целом уравнение множественной регрессии является очень значимым (см. материал ниже для обсуждения проверки статистической значимости). Таким образом, учитывая независимые переменные, вы можете «предсказать» бедность лучше, чем можно было бы ожидать по чистой случайности.



Multiple Regression Results		
Dependent: PT_POOR	Multiple R = .88215345	F = 13.44909
	R² = .77819471	df = 6,23
No. of cases: 30	adjusted R² = .72033245	p = .000002
Standard error of estimate: 3.398607071		
Intercept: 31.266044090	Std. Error: 13.26511	t(23) = 2.3570 p = .0273
POP_CHNG beta=-.63	N_EMPLD beta=.238	TAX_RATE beta=.039
PT_PHONE beta=-.13	PT_RURAL beta=.619	AGE beta=-.19
(significant betas are highlighted)		

??? Что такое «статистическая значимость» (p-уровень). Статистическая значимость результата — это оценочная мера степени, в которой он является «верным» (в смысле «представителем населения»). С технической точки зрения, значение p-уровня представляет собой убывающий показатель надежности результата. Чем выше p-уровень, тем меньше мы можем полагать, что наблюдаемая связь между переменными в выборке является надежным индикатором связи между соответствующими переменными в совокупности. В частности, p-уровень представляет собой вероятность ошибки, связанной с принятием нашего наблюдаемого результата как достоверного, то есть как «представителя совокупности». Например, p-уровень 0,05 (т. е. 1/20) указывает на то, что существует 5% вероятность того, что связь между переменными, обнаруженными в нашей выборке, является «случайной». Другими словами, если предположить, что в популяции между этими переменными нет никакой связи, и мы повторяем опыты, подобные нашему, один за другим, мы могли бы ожидать, что примерно в каждом 20 повторениях опыта будет одна, в которой связь между рассматриваемыми переменными была бы равна или сильнее, чем в нашей. Во многих областях исследований p-уровень 0,05 обычно рассматривается как «границно приемлемый» уровень ошибки.

Коэффициенты регрессии. Чтобы узнать, какая из независимых переменных больше всего способствует прогнозированию бедности, изучите коэффициенты регрессии (или B). Нажмите кнопку «Сводка: результаты регрессии» Summary: Regression results на вкладке «Быстрая», чтобы отобразить электронную таблицу с этими коэффициентами.

Data: Regression Summary for Dependent Variable: PT_POOR (Povert...						
Regression Summary for Dependent Variable: PT_POOR (Povert						
R= .88215345 R²= .77819471 Adjusted R²= .72033245						
F(6,23)=13.449 p<.00000 Std.Error of estimate: 3.3986						
N=30	b*	Std.Err. of b*	b	Std.Err. of b	t(23)	p-value
Intercept			31.26604	13.26511	2.35701	0.027309
POP_CHNG	-0.630788	0.129413	-0.39234	0.08049	-4.87421	0.000064
N_EMPLD	0.238314	0.140987	0.00075	0.00044	1.69033	0.104476
TAX_RATE	0.038799	0.100611	1.23012	3.18985	0.38563	0.703311
PT_PHONE	-0.129627	0.203294	-0.08325	0.13056	-0.63763	0.530012
PT_RURAL	0.618746	0.231173	0.16554	0.06185	2.67655	0.013476
AGE	-0.188205	0.114652	-0.41926	0.25541	-1.64153	0.114292

В этой электронной таблице показаны стандартизированные коэффициенты регрессии (b*, см. пояснения ниже) и необработанные коэффициенты регрессии (b). Величина этих бета-коэффициентов позволяет сравнивать относительный вклад каждой независимой переменной в прогноз зависимой переменной. Как видно из показанной выше электронной таблицы, переменные POP_CHNG, PT_RURAL и N_EMPLD являются наиболее важными предикторами бедности; из них только первые две переменные являются статистически значимыми. Коэффициент регрессии для POP_CHNG отрицательный; чем меньше прирост населения, тем больше число семей, живущих за чертой бедности в соответствующем округе. Вес регрессии для PT_RURAL положительный; чем больше процент сельского населения, тем выше уровень бедности.

??? Стандартизированный коэффициент регрессии рассчитывается по формуле

$$\beta_j = b_j \frac{\delta_{xj}}{\delta_y}$$

β_j – стандартизированный коэффициент при факторе x_j . Определяет силу влияние вариации x_j на вариацию результативного признака y при отвлечении от сопутствующего влияния вариаций других факторов, входящих в уравнение регрессии. Т.к. β_j сравнимы между собой, то по величине данных коэффициентов можно ранжировать факторы по силе их воздействия на результат. Смысл стандартизованных коэффициентов β_j позволяет использовать их при отсеивании факторов, т.е. из модели исключаются факторы с наименьшим значением β_j .

Частичные корреляции. Другой способ посмотреть на уникальный вклад каждой независимой переменной в прогноз зависимой переменной — это вычислить частичные и получастичные корреляции (нажмите кнопку «Частичные корреляции» на вкладке «Дополнительно» в диалоговом окне «Результаты»). Частичные корреляции — это корреляции между соответствующей независимой переменной, скорректированной всеми другими переменными, и зависимой переменной, скорректированной всеми другими переменными. Таким образом, это корреляция между остатками после корректировки всех независимых переменных. Частичная корреляция представляет собой уникальный вклад соответствующей независимой переменной в предсказание зависимой переменной.

Data: Variables currently in the Equation; DV: PT_POOR (Poverty)*							
Variable	b* in	Partial Cor.	Semipart Cor.	Tolerance	R-square	t(23)	p-value
POP_CHNG	-0.630788	-0.712815	-0.478659	0.575818	0.424182	-4.87421	0.000064
N_EMPLD	0.238314	0.332414	0.165994	0.485159	0.514841	1.69033	0.104476
TAX_RATE	0.038799	0.080152	0.037870	0.952684	0.047316	0.38563	0.703311
PT_PHONE	-0.129627	-0.131795	-0.062617	0.233342	0.766658	-0.63763	0.530012
PT_RURAL	0.618746	0.487340	0.262844	0.180456	0.819544	2.67655	0.013476
AGE	-0.188205	-0.323838	-0.161202	0.733638	0.266362	-1.64153	0.114292

Получастичная корреляция — это корреляция соответствующей независимой переменной, скорректированной всеми другими переменными, с необработанной (нескорректированной) зависимой переменной. Таким образом, получастичная корреляция представляет собой корреляцию остатков для соответствующей независимой переменной после корректировки всех других переменных и нескорректированных необработанных оценок для зависимой переменной. Иными словами, квадрат получастной корреляции является индикатором процента общей дисперсии, однозначно учитываемой соответствующей независимой переменной, в то время как квадрат частной корреляции является индикатором процента остаточной дисперсии, учитываемой после корректировки зависимой переменной для все остальные независимые переменные.

В этом примере частичная и получастичная корреляции относительно схожи. Однако иногда их величина может сильно различаться (получастичная корреляция всегда ниже). Если получастичная корреляция очень мала, а частичная корреляция относительно велика, то соответствующая переменная может предсказать уникальный «фрагмент» изменчивости в зависимой переменной (который не учитывается в других переменных). Однако с точки зрения практической значимости этот фрагмент может быть крошечным и представлять лишь очень небольшую долю общей изменчивости.

Анализ остатков. После подгонки уравнения регрессии всегда следует проверять предсказанные и остаточные оценки. Например, крайние выбросы могут серьезно исказить результаты и привести к ошибочным выводам. В диалоговом окне «Результаты множественной регрессии» — вкладка «Остатки/допущения/прогноз» нажмите кнопку «Выполнить анализ невязок», чтобы перейти к диалоговому окну «Анализ невязок».

График случайных остатков. Как правило, вы должны, по крайней мере, изучить образец необработанных или стандартизированных остатков, чтобы выявить какие-либо экстремальные выбросы. В этом примере выберите вкладку «Остатки» и нажмите кнопку «Суммарный график остатков» (*Casewise plot of residuals*); по умолчанию необработанные остатки будут «нанесены» на график по случаям (электронная таблица); однако вы также можете выбрать другую остаточную статистику в группе «Тип остатка».

Data: Raw Residual (poverty*)						Raw Residual (poverty.sta)			
						Dependent variable: PT_POOR			
Case name	Raw Residuals					Observed Value	Predicted Value	Residual	Standard Pred. v.
	-3s	.	0	.	+3s				
Benton	.	.	*	.	.	19.00000	19.04284	-0.04284	-0.69977
Cannon	.	.	*	.	.	26.20000	30.66326	-4.46326	1.34996
Carrol	.	.	*	.	.	18.10000	20.07600	-1.97600	-0.51753
Cheatheam	.	.	*	.	.	15.40000	15.82975	-0.42975	-1.26653
Cumberland	*	29.00000	24.72018	4.27982	0.30166
DeKalb	.	.	*	.	.	21.60000	24.16319	-2.56319	0.20341
Dyer	.	.	.	*	.	21.90000	21.20206	0.69794	-0.31890
Gibson	.	.	*	.	.	18.90000	18.91778	-0.01778	-0.72183

Шкала, используемая на графике по случаям в крайнем левом столбце, представлена в терминах сигмы, то есть стандартного отклонения остатков. Если один или несколько случаев выходят за пределы ± 3 -кратного сигма, вероятно, следует исключить соответствующие случаи (что легко сделать с помощью условий отбора случаев) и запустить анализ, чтобы убедиться, что эти выбросы не искажают ключевые результаты.

График случайных выбросов. Быстрый способ определить выбросы — нажать кнопку График выбросов по случаям (*Casewise plot of outliers*) на вкладке «Выбросы». Вы можете либо построить все стандартные невязки, выходящие за пределы ± 2 сигма, либо построить 100 наиболее экстремальных случаев, как указано в поле Тип выброса. При выборе кнопки опции Стандартная невязка (или *Standard residual*) ($> 2 * \text{сигма}$) выбросы в текущем примере обнаружены не будут.

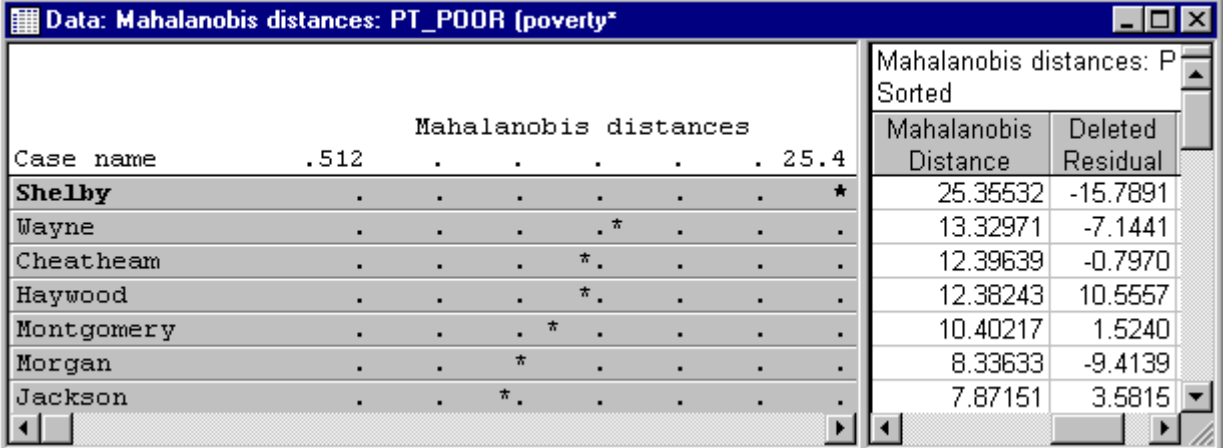
Расстояния Махаланобиса (Mahalanobis distances). Большинство учебников по статистике посвящают некоторое обсуждение вопросу о выбросах и остатках зависимой переменной. Однако роль выбросов в списке независимых переменных часто упускается из виду. На стороне независимой переменной у вас есть список переменных, которые участвуют с разными весами (коэффициенты регрессии) в прогнозировании зависимой переменной. Вы можете думать о независимых переменных как об определении многомерного пространства, в котором может быть расположено каждое наблюдение.

Например, если у вас есть две независимые переменные с одинаковыми коэффициентами регрессии, вы можете построить диаграмму рассеяния этих двух переменных и поместить каждое наблюдение на этот график.

Затем вы можете построить одну точку для среднего значения обеих переменных и вычислить расстояния каждого наблюдения от этого среднего (теперь называемого

центроидом) в двумерном пространстве; это концептуальная идея вычисления расстояний Махаланобиса.

Теперь посмотрите на эти расстояния (отсортированные по размеру), чтобы определить крайние случаи на стороне независимой переменной. Выберите кнопку опции Расстояния Махаланобиса в поле Тип выбросов, а затем нажмите кнопку График выбросов по случаям. Результирующий график (таблица результатов) покажет расстояния Махаланобиса, отсортированные в порядке убывания.

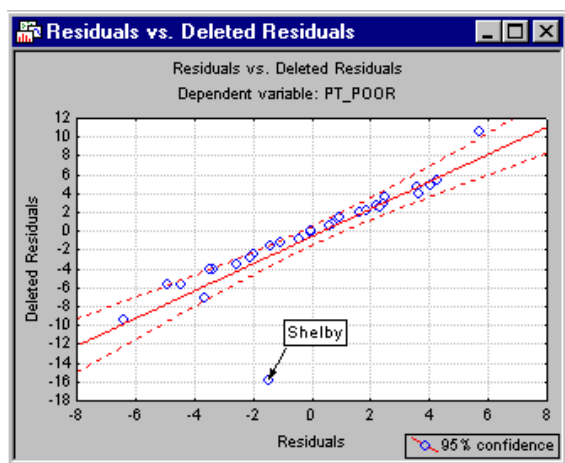


Case name	Mahalanobis distances
	.512 25.4
Shelby *
Wayne * . .
Cheatheam	. . . * . . .
Haywood	. . . * . . .
Montgomery	. . . * . . .
Morgan	. . *
Jackson	. . *

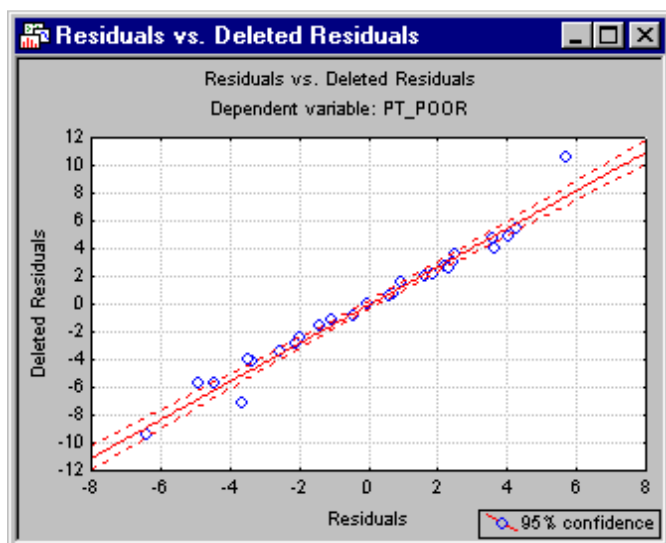
Mahalanobis Distance	Deleted Residual
25.35532	-15.7891
13.32971	-7.1441
12.39639	-0.7970
12.38243	10.5557
10.40217	1.5240
8.33633	-9.4139
7.87151	3.5815

Обратите внимание, что округ Шелби (Shelby, в первой строке) выглядит несколько экстремально по сравнению с другими округами на графике. Если вы посмотрите на необработанные данные, вы обнаружите, что округ Шелби действительно является самым большим округом в файле данных, в котором гораздо больше людей занято в сельском хозяйстве (переменная N_EMPLD) и т. д. Вероятно, было бы разумно выразить эти числа в процентах, а не в абсолютных числах, и в этом случае расстояние Махаланобиса округа Шелби от других округов в выборке, вероятно, не было бы таким большим. Однако в нынешнем виде округ Шелби явно выделяется.

Удаленные остатки. Другой очень важной статистикой, позволяющей оценить серьезность проблемы выброса, является удаленный остаток. Это стандартизированный остаток для соответствующего случая, который можно было бы получить, если бы случай был исключен из анализа. Помните, что процедура множественной регрессии соответствует прямой линии, чтобы выразить отношения между зависимыми и независимыми переменными. Если один случай явно является выбросом (как округ Шелби в этих данных), то существует тенденция к тому, чтобы линия регрессии «вытягивалась» этим выбросом, чтобы максимально учесть его. В результате, если бы соответствующий случай был исключен, получилась бы совсем другая линия (и коэффициенты B). Следовательно, если удаленный остаток сильно отличается от стандартизированного остатка, у вас есть основания полагать, что регрессионный анализ серьезно искажен соответствующим случаем. В этом примере удаленный остаток для округа Шелби является выбросом, который серьезно влияет на анализ. Вы можете построить график остатков по отношению к удаленным остаткам с помощью кнопки Остатки по сравнению с удаленными остатками на вкладке Диаграммы рассеяния, которая создаст диаграмму рассеивания этих значений. Диаграмма рассеяния ниже ясно показывает выброс.

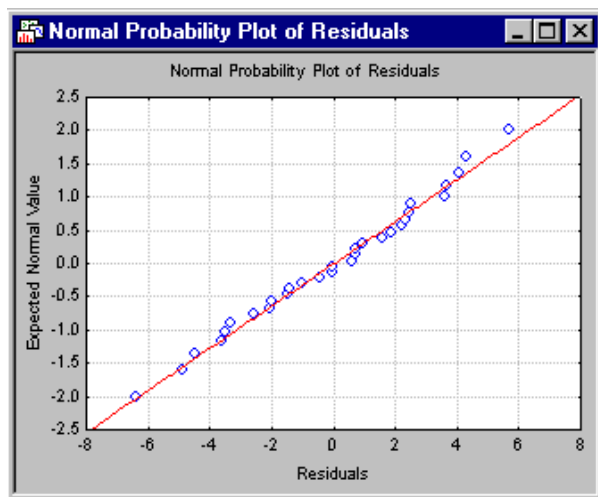


STATISTICA предоставляет интерактивный инструмент для удаления выбросов (Инструмент «Кисть», Brushing Tool), позволяющий экспериментировать с удалением выбросов и мгновенно видеть их влияние на линию регрессии. Щелкните правой кнопкой мыши график и выберите Show Brushing в контекстном меню. Когда инструмент активен, курсор принимает форму перекрестия, а рядом с графиком отображается диалоговое окно Brushing. Вы можете (временнo) в интерактивном режиме исключить отдельные точки данных из графика, установив 1) флажок «Автоматически применять» и 2) кнопку выбора «Выкл.» в поле «Действие»; затем щелкните перекрестием точку, которую нужно удалить. Щелчок по точке автоматически удаляет ее (временнo) с графика.



Обратите внимание, что вы можете «вернуть» удаленные точки, нажав кнопку «Сбросить все» в диалоговом окне «Brushing».

Графики нормальной вероятности. В диалоговом окне «Анализ остатка» доступно множество дополнительных графиков. Большинство из них более или менее прямолинейны в своей интерпретации; однако графики нормальной вероятности будут здесь прокомментированы. Как упоминалось ранее, множественная линейная регрессия предполагает линейные отношения между переменными в уравнении и нормальное распределение остатков. Если эти предположения нарушаются, ваш окончательный вывод может быть неточным. График нормальной вероятности остатков покажет вам, имели ли место грубые нарушения допущений. Нажмите кнопку Нормальный график остатков на вкладке Графики вероятностей, чтобы отобразить этот график.



Этот график строится следующим образом. Сначала остатки упорядочены по рангу. Из этих рангов можно вычислить значения z (т. е. стандартные значения нормального распределения) на основе предположения, что данные поступают из нормального распределения. Эти значения z нанесены по оси y на графике.

Если наблюдаемые остатки (отложенные на оси x) нормально распределены, то все значения должны падать на прямую линию на графике; на этом графике все точки очень близко следуют за линией. Если остатки не распределены нормально, то они будут отклоняться от прямой. Выбросы также могут стать очевидными на этом графике.

Если в целом наблюдается несоответствие и кажется, что данные образуют четкую структуру (например, S-образную форму) вокруг линии, тогда зависимую переменную, возможно, придется каким-то образом преобразовать (например, логарифмическое преобразование для «вытягивания» в хвосте распределения. Обсуждение таких техник выходит за рамки этого примера.

6 Контрольные вопросы

1. Дать определение регрессионного анализа (regression analysis)
2. Дать определение уравнение регрессии
3. Что такое результирующая переменная и предикторы в регрессионном анализе? переменными.
4. Дать математическое определение регрессии
5. Какие существуют модели уравнений регрессии
6. Дать уравнение множественной линейной регрессии
7. Каким образом применяется метод наименьших квадратов в регрессионном анализе?
8. Какой функционал минимизируется при применении метода наименьших квадратов в регрессионном анализе?
9. Дать понятие функции невязки

10. Для чего применяются диаграммы рассеяния в регрессионном анализе?
11. Что такое «статистическая значимость» (p-уровень)? Показать в своих результатах
12. Записать полученное по результатам работы уравнение регрессии
13. Дать определение стандартизированного коэффициента регрессии
14. Дать определение частичным корреляциям
15. Что из себя представляют остатки регрессионной модели?
16. С какой целью осуществляется анализ остатков регрессионной модели?
17. Дать понятие случайных выбросов
18. Расстояния Махаланобиса (Mahalanobis distances) – дать определение и указать область применения в регрессионном анализе
19. Что позволяет продемонстрировать опция «Удаленные остатки» а пакете Статистика?