

ЛАБОРАТОРНАЯ РАБОТА №8 РЕГРЕССИЯ

ПРОСТАЯ РЕГРЕССИЯ

Процедура простой регрессии заключается в нахождении аналитического выражения для связи двух переменных. Модели простой регрессии, предусмотренные в *Statgraphics*, представлены в таблице

Тип модели	Связь
Линейная	$Y=a+b*X$
Экспоненциальная	$Y=\exp(a+b*X)$
Обратная по Y	$Y=1/(a+b*X)$
Обратная по X	$Y=a+b/X$
Дважды обратная	$Y=1/(a+b/X)$
Логарифм по X	$Y=a+\ln(X)$
Мультипликативная	$Y=a*X^b$
Квадратный корень по X	$Y=a+b*\sqrt{X}$
Квадратичная	$Y=(a+b*X)^2$
S- кривая	$Y=\exp(a+b/X)$
Логистическая	

Американским астрономом Хабблом в 1929 году было обнаружено, что галактики удаляются от Земли тем быстрее, чем дальше они расположены. Также им было установлено, что скорость удаления пропорциональна расстоянию. Коэффициент этой пропорциональности получил название **постоянной Хаббла**. О его точном значении в астрономии продолжается дискуссия, хотя сама идея линейной зависимости признана безусловно. В настоящее время указанное явление истолковывается как свидетельство расширения вселенной.

Задача. С помощью модели простой регрессии оценим постоянную в законе Хаббла.


Данные, которые будем подвергать анализу, представляют собой расстояния от Земли (в миллионах световых лет) и скорости удаления (в сотнях миль в секунду) 11 галактик.

Вам представлены следующие данные:

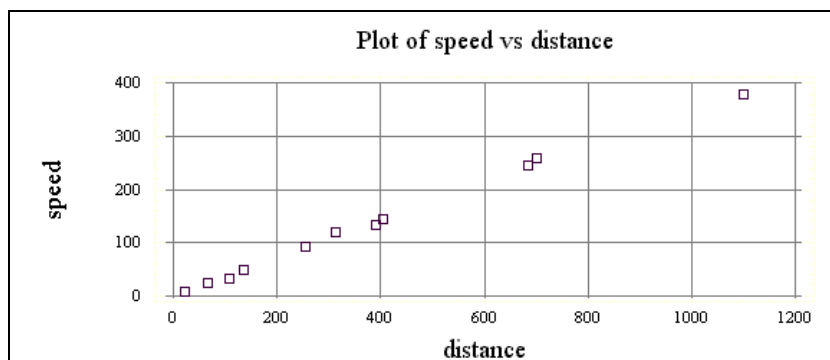
Условное название галактики	Расстояние	Скорость
Дева	22	7,5
Пегас	68	24
Персей	108	32
Волосы Вероники	137	47
Большая Медведица 1	255	93
Лев	315	120
Северная корона	390	134
Близнецы	405	144
Волопас	685	245
Большая Медведица 2	700	260
Гидра	1100	380

Введите их в электронную таблицу *Statgraphics*. Должно получиться следующее:

	distance	speed
1	22	7,5
2	68	24
3	108	32
4	137	47
5	255	93
6	315	120
7	390	134
8	405	144
9	685	245
10	700	260
11	1100	380

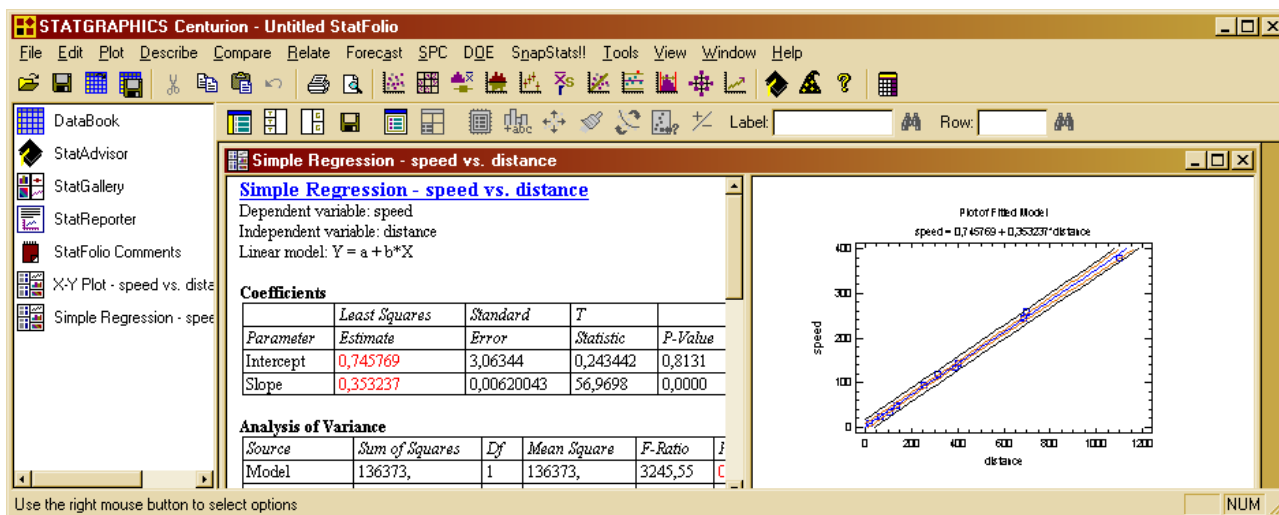
Сохраните данные. Посмотрим теперь графическое представление введенной таблицы. Для быстроты и простоты нажмите кнопку  **ScatterPlot** на панели инструментов. Появится окно **X-Y Plot**, в котором щелкните по **speed**, щелкните по кнопке со стрелкой в поле **Y**, затем выберите **distance**, щелкните по кнопке со стрелкой в поле **X**.

Нажмите ОК. Появится график:



Видно, что точки достаточно хорошо укладываются на прямую, поэтому стоит попробовать воспользоваться линейной регрессией. Кстати, точки можно соединить отрезками прямых. Для этого надо щелкнуть по графику правой кнопкой, в контекстном меню выбрать **Pane Options**, в раскрывшемся диалоговом окне включить **Lines**.

В строке меню выберите **Relate**, в раскрывшемся меню выберите **One Factor**, затем **Simple Regression**. Раскроется окно **Simple Regression**, щелкните по **speed**, щелкните по кнопке со стрелкой в поле **Y**, затем выберите **distance**, щелкните по кнопке со стрелкой в поле **X**. Нажмите ОК. Раскроется окно — рабочее поле процедуры простой регрессии со статистической сводкой применительно к линейной модели.

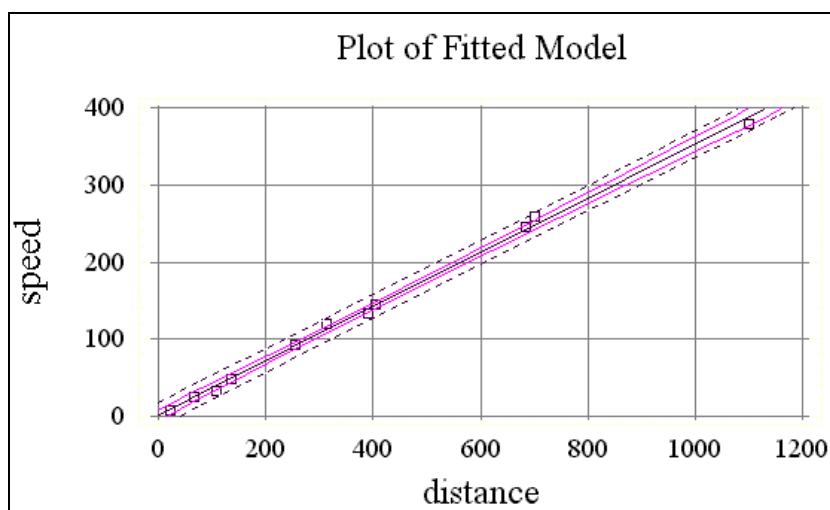


Щелкните дважды по окну с анализом. Обратите внимание, коэффициент корреляции (*Correlation Coefficient*) равен 0,998616 — это говорит о том, что построена очень хорошая модель, сильно коррелирующая с экспериментальными наблюдениями. Об этом же свидетельствуют результаты дисперсионного ана-

лиза модели (*Analysis of Variance*): достаточно высокий коэффициент детерминации *R-squared* (R-квадрат), низкое значение *p-value*.

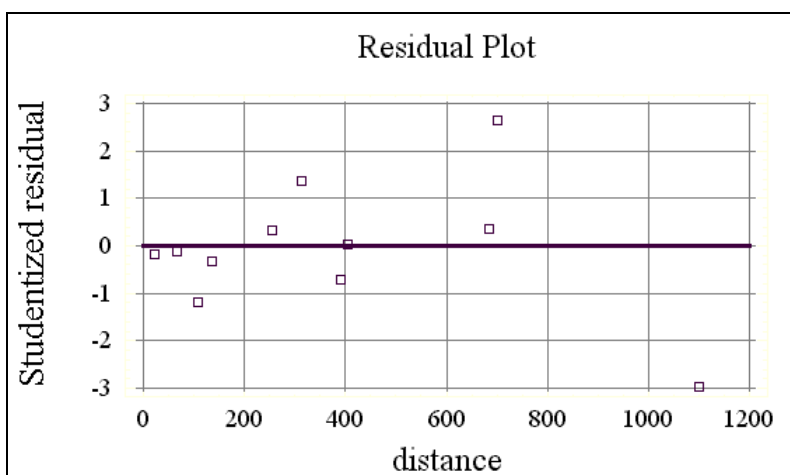
Угол наклона (*Slope*) равен 0,353237, это и есть постоянная Хаббла. Значение *p-value* для этого коэффициента, меньший 0,05, говорит о статистической значимости этого коэффициента. К сожалению, второй коэффициент (*intersept*) не является статистически значимым. **Стандартная ошибка оценки** (*Standard Error of Est*) равна 6,48216.

Посмотрим теперь на графическое представление результатов. Выборочная линия регрессии лежит в центре двух «трубок». Внутренняя трубка — это доверительный интервал для оцениваемой нами линейной функции регрессии, внешняя — это доверительный интервал для самих наблюдений.



Можно изменить доверительные вероятности для этих интервалов. Для этого щелкните правой кнопкой по графику, выберите *Pane Options*, перед вами раскроется окно *Plot of Fitted Model Options*, выберите в нем *Confidence Limits*. Также в этом окне можно менять *Confidence Level*, по умолчанию стоит 95 %.

Построим теперь **график остатков**. Для этого нажмите кнопку **Graphs**, в раскрывшемся окне выберите *Residual versus X* (график остатков). Перед вами раскроется окно.



Этот график заставляет задуматься: напрашивается наличие какой-то периодической компоненты в анализируемых измерениях. Является ли она следствием использования технологии измерений или имеется другая причина — есть повод для поиска объяснений.

В целом подтверждается гипотеза Хаббла о линейной зависимости скорости удаления звезд от их расстояния от Земли, а также получено значение постоянной Хаббла, хорошо согласующееся с известными данными.

На самом деле, мы с самого начала предполагали линейную зависимость. Поскольку *Statgraphics* позволяет использовать не только линейную модель, попробуем воспользоваться другими возможными моделями. Нажмите кнопку **Tables**, в раскрывшемся окне выберите **Comparison of Alternative Models** (сравнение альтернативных моделей). Нажмите ОК.

Comparison of Alternative Models		
Model	Correlation	R-Squared
Double reciprocal	0,9993	99,86%
Double square root	0,9988	99,75%
Multiplicative	0,9987	99,74%
Linear	0,9986	99,72%
Double squared	0,9981	99,61%
Square root-X	0,9780	95,65%
Square root-Y	0,9691	93,91%
Squared-Y	0,9637	92,88%
Square root-Y logarithmic-X	0,9613	92,41%
Logarithmic-Y square root-X	0,9586	91,89%
Squared-X	0,9428	88,89%
Logarithmic-X	0,8834	78,04%
Squared-Y square root-X	0,8822	77,83%
Exponential	0,8727	76,16%
Reciprocal-Y logarithmic-X	-0,8709	75,85%
S-curve model	-0,8620	74,30%
Square root-Y squared-X	0,8550	73,11%
Squared-Y logarithmic-X	0,7338	53,84%
Logarithmic-Y squared-X	0,7087	50,22%
Square root-Y reciprocal-X	-0,7082	50,15%
Reciprocal-X	-0,5711	32,62%

Перед вами раскрылась таблица, в которой представлены результаты анализа для всех типов зависимостей Y от X , упорядоченные по коэффициенту корреляции с экспериментальными наблюдениями. Оказывается, что линейная модель занимает лишь третье место по качеству аппроксимации экспериментальных наблюдений. На первом же месте модель с дважды обратным преобразованием. Однако их преимущество не очень значительно, а линейная модель привлекает своей простотой.

Simple Regression Options

Type of Model

☒ Linear
☐ Square Root-Y
☐ Exponential
☐ Reciprocal-Y
☐ Squared-Y
☐ Square Root-X
☐ Double Square Root
☐ Log-Y Square Root-X
☐ Reciprocal-Y Square Root-X
☐ Squared-Y Square Root-X
☐ Logarithmic-X
☐ Square Root-Y Log-X
☐ Multiplicative
☐ Reciprocal-Y Log-X
☐ Squared-Y Log-X
☐ Reciprocal-X
☐ Square Root-Y Reciprocal-X
☐ S-Curve
☐ Double Reciprocal
☐ Squared-Y Reciprocal-X
☐ Squared-X
☐ Square Root-Y Squared-X
☐ Log-Y Squared-X
☐ Reciprocal-Y Squared-X
☐ Double Squared
☐ Logistic
☐ Log Probit

Alternative Fit

☒ None (least squares only)
☐ Minimize absolute deviations
☐ Use medians of 3 groups

OK

Cancel

Help

Посмотрим внимательнее на результаты с **мультипликативной** моделью. Для этого щелкните правой кнопкой в окне сравнения моделей, выберите **Analysis Options**, перед Вами раскроется окно, в котором необходимо выбрать нужный Вам тип модели (*multiplicative*). Нажмите ОК, дважды щелкните по таблице результатов, раскроется таблица с результатами регрессионного анализа для данной модели. Проанализируйте полученные результаты.

Проанализируйте самостоятельно регрессионную модель дважды обратного преобразования. Результаты покажите преподавателю.

МНОЖЕСТВЕННАЯ РЕГРЕССИЯ

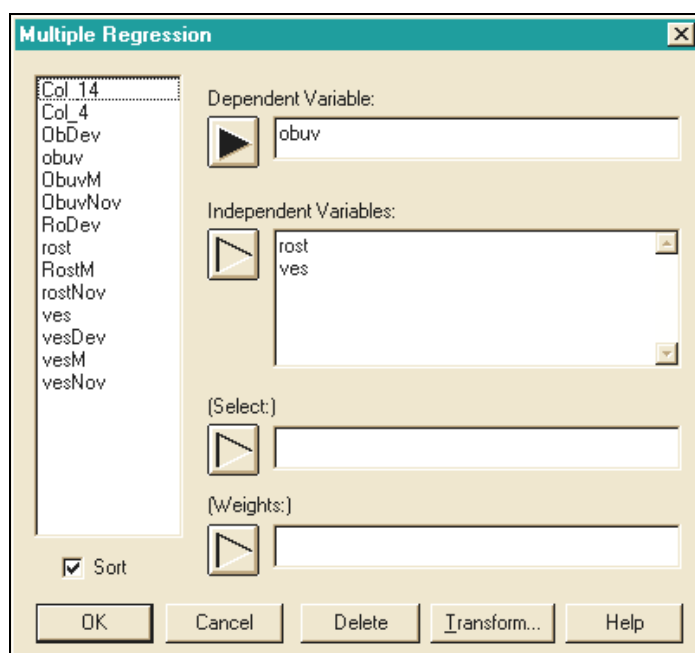
Предметом множественного регрессионного анализа является установление статистической зависимости среднего значения одной случайной величины Y от нескольких других величин X_1, X_2, \dots, X_n . Эта статистическая зависимость находит свое выражение в уравнении:

$$Y = a_0 + a_1 X_1 + a_2 X_2 + \dots + a_n X_n, \text{ где } a_i (i = 0, 1, \dots, n) \text{ — искомые параметры}$$

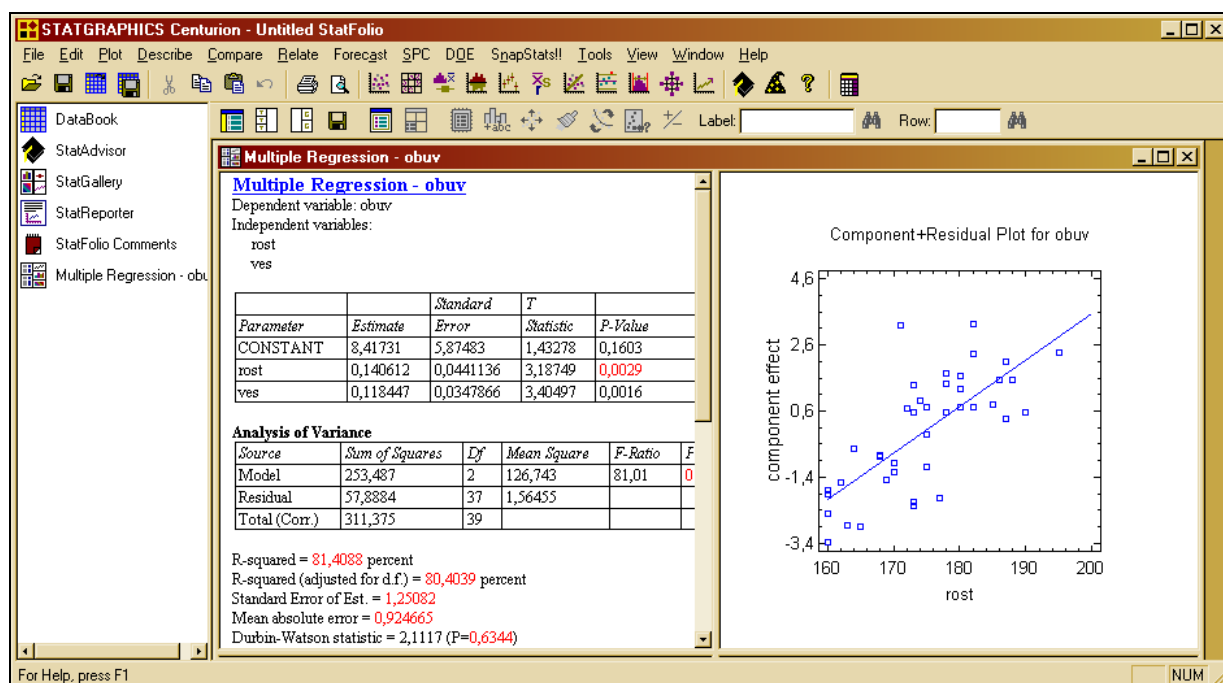
Задача. В файле *Rost_Razmer* приведены данные о 40 студентах (группы 401 – 403) — их рост, размер обуви, вес. Провести корреляционный анализ, найти зависимость размера обуви от остальных величин.

Самостоятельно проведите корреляционный анализ, затем найдите уравнения регрессионной зависимости размера обуви от роста и от веса. Результаты покажите преподавателю.

В строке меню выберите **Relate**, в раскрывшемся меню выберите **Multiple Factor**, затем **Multiple Regression** (множественная регрессия). Раскроется окно, в котором выберите *Obuv*, щелкните по стрелке в поле **Dependent Variable** (зависимая переменная), затем выберите *rost*, щелкните по стрелке в поле **Independent Variable** (независимая переменная).



Аналогично поступите с *ves*. Нажмите кнопку ОК. Перед вами раскроется окно анализа множественной регрессии.

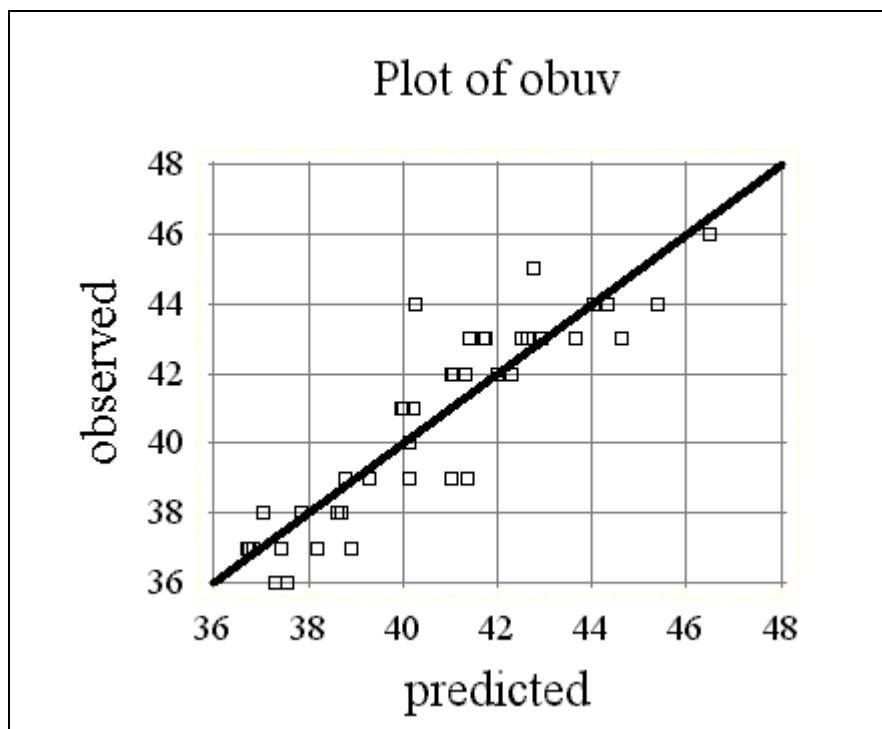


Получили уравнение регрессии: $Obuv = 8,4173 + 0,1406 * rost + 0,1184 * ves$.

В соответствии со значением статистики *R-squared* видим, что модель отражает 81,4 % изменчивости переменной *Obuv*, а скорректированный R-квадрат с учетом степеней свободы (что является более подходящим для сравнения моделей с разными количествами переменных) составляет 80,4 %. Стандартная ошибка равна 0,92, и ее можно использовать в задании границ предсказания для

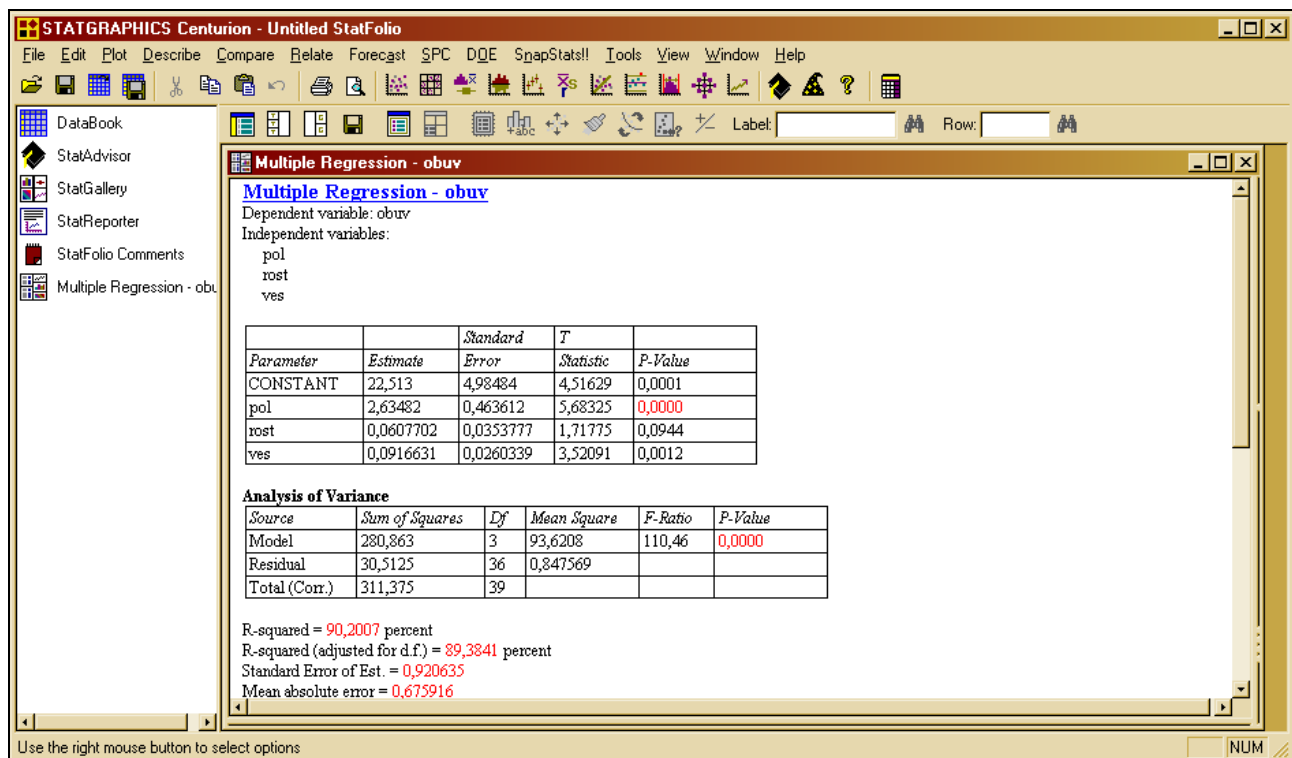
новых наблюдений. Средняя абсолютная ошибка, представляющая собой среднюю величину остатков, составляет 0,92. К сожалению, константа не является статистически значимой.

Для множественной регрессии графическое представление полученных результатов на плоскости можно получить следующим образом: щелкните по кнопке **Graphs**, в появившемся окне выберите **Observed versus Predicted** (наблюдения — предсказания). Раскроется следующее окно.



Исходя из этого графика заключаем, что полученная модель заслуживает доверия и зафиксированные взаимоотношения могут быть подвергнуты дальнейшей содержательной интерпретации.

Попробуем улучшить результат. Очевидно, что размер обуви зависит от пола студента. Введем фиктивную переменную (pol): 0 — девочка; 1 — мальчик.



Видно, что все коэффициенты статистически значимы при уровне значимости 0,1. Убедитесь в том, что все характеристики улучшились.

ЗАДАНИЯ

Задача 1. Данные о расходе электроэнергии (кВт/ч) на изготовление одной тонны цемента (Y) в зависимости от объема выпуска (X) продукции (тыс.т) цементными заводами приводится в таблице

Выпуск продукции	5	10	15	20	25	30
Расход электроэнергии Y	10,0	8,2	7,3	6,3	6,4	5,2

Построить уравнение регрессии $Y = a_0 + a_1(1/x)$.

1. Приемлема ли эта модель? Почему?
2. Является ли эта модель наилучшей?

Задача 2. Для про производство электроэнергии и рекорды по прыжкам в высоту из предыдущей лабораторной работы (файл E_H) найти $\tilde{E} = a_0 + a_1T$, $\tilde{H} = b_0 + b_1T$, (здесь переменная T это Year), сохранить остатки $e_E = E - \tilde{E}$, $e_H = H - \tilde{H}$. Сравните коэффициенты корреляции остатков и частный коэффициент корреляции переменных E и H при фиксированной переменной Year.

Задача 3. В таблице представлены данные о темпах прироста (%) следующих макроэкономических показателей $n=10$ развитых стран мира за 1992 г.: ВВП — $x^{(1)}$, промышленного производства — $x^{(2)}$, индекса цен — $x^{(3)}$.

Страны	$x^{(1)}$	$x^{(2)}$	$x^{(3)}$
Япония	3,5	4,3	2,1
США	3,1	4,6	3,9
Германия	2,2	2,0	3,4
Франция	2,7	3,1	2,9
Италия	2,7	3,0	5,6
Великобритания	1,6	1,4	4,0
Канада	3,1	3,4	3,0
Австралия	1,8	2,6	3,4
Бельгия	2,3	2,6	3,4
Нидерланды	2,3	2,4	3,5

Приняв за объясняемую величину (y) показатель $x^{(1)}$, а за объясняющую (x) — $x^{(2)}$, построить уравнения регрессии:

1. $y = a_0 + a_1 x$
2. $y = a * x^b$.

На какой модели вы остановитесь? Улучшится ли результат, если к независимым переменным добавить $x^{(3)}$?

Задача 4. В файле **Питер.sf** выясните, от чего зависит цена квартиры и цена одного квадратного метра. Попробуйте подобрать приемлемую регрессионную модель для цены квартиры (и цены одного квадратного метра).

ВОПРОСЫ

1. В каких случаях выборочную функцию регрессии следует искать в виде линейной функции?
2. Какой метод используется для нахождения коэффициентов линейной функции регрессии?
3. Что характеризует коэффициент детерминации R^2 ?
4. Что такое остаточная дисперсия? Что она характеризует?

5. Что можно сказать про остаточную дисперсию, если выборочный коэффициент корреляции близок к 1?
6. Для чего используется критерий Дарбина-Уотсона?
7. Что следует проверить при анализе остатков?