Глава 16

Тонкая настройка регрессионного прогноза

В этой главе…

- Использование множественной регрессии
- Вывод на график линии тренда регрессии
- Анализ прогнозов, выполненных методом регрессии

В других главах, в частности в главе 11, мы использовали регрессию для прогнозирования одной переменной (в частности, объема продаж) по переменной предиката (номеру периода времени или количеству торговых агентов). Этот вид анализа чаще всего называют *простой регрессией*. В то же время возможно прогнозирование одной переменной по нескольким предикатам, и в этом подходе есть свои преимущества. К примеру, вы можете попытаться создать прогноз по значениям как периода времени, так и количества торговых представителей. Этот подход называется *множественной регрессией*, и эта глава посвящена его реализации в программе Excel.

Одним из самых замечательных аспектов диаграмм в Excel является возможность отображения линии тренда, с помощью которой всего за один шаг можно отобразить взаимосвязь прогнозируемой переменной с предикатом. Линия тренда может быть как линейной, так и нелинейной, также она может быть реализацией скользящего среднего. Линия тренда визуально информирует пользователя о направлении и силе взаимосвязи предикатов с переменной прогнозирования. При желании вы можете также отобразить на диаграмме числовое значение “R-квадрат” и само уравнение регрессии.

Довольно заманчиво, не прилагая никаких усилий, взять числа, которые вычислены компьютером, даже не обработав их. Однако этого делать нельзя: вас ожидает множество ловушек. Вы сможете избежать неприятностей, если оцените полученный прогноз, прежде чем полностью ему доверитесь, и в этой оценке вам смогут помочь некоторые инструменты Excel.

Выполнение множественной регрессии

Множественная регрессия является способом использования более одной переменной предиката для предсказания другой переменной, такой как доходы от продаж. Это делает прогноз более точным, однако при этом нужно знать, *как* заставить компьютер (в данном случае персональный с установленной на нем программой Excel) выполнить то, что называют множественной регрессией. И не только это: вам нужно еще знать, как *интерпретировать* полученные результаты. В этой главе вы научитесь и тому, и другому.

Использование более одного предиката

Как одновременно использовать несколько предикатов для предсказания одной переменной? Именно здесь на помощь приходит множественная регрессия. Она объединяет участие нескольких переменных в формировании одной.

Для примера предположим, что вас интересует прогноз веса человека в диапазоне от младенческого возраста до совершеннолетия. В качестве исходного материала у вас под рукой могут быть данные о весе, росте и весе выборочной группы людей (например, тридцати человек). Если организовать эти данные правильно на рабочем листе и направить внимание программы Excel на эти три переменные, ее функция регрессии сделает следующее.

- ✓ Скомбинирует влияние двух переменных (роста и возраста) на формирование одной (веса).
- ✓ Вычислит наиболее вероятные значения этой новой переменной для имеющегося набора исходных данных о предикатах.

К примеру, функция регрессии может вычислить новую переменную следующим образом:

```
Новая_переменная=-8,02+(1,58×Возраст)+(0,35×Рост)
```

Функция регрессии вычислит коэффициенты для предикатов возраста и роста (`1,58` и `0,35`) так, чтобы новая переменная имела максимальную корреляцию с весом для имеющегося набора данных. Функция регрессии пакета анализа не показывает значения этой новой переменной. В то же время она предоставляет данные, позволяющие легко вычислить их собственноручно: коэффициенты предикатов и постоянную, формирующие *уравнение регрессии*. Вы можете использовать и встроенную функцию `ТЕНДЕНЦИЯ` (`TREND`) самой программы Excel, которая отображает значения новой переменной на экране.

Новая переменная на самом деле является прогнозом значения веса. При этом для ее получения коэффициенты регрессии умножаются на соответствующие им предикаты, после чего результаты складываются между собой. На рис. 16.1 этот процесс показан более детально.

На нем проиллюстрирована рассматриваемая в качестве примера в этом разделе задача: прогнозирование веса по росту и возрасту. (Я понимаю, прогнозирование веса вас, скорее всего, не интересует вовсе. Однако этот пример упростит предстоящую дискуссию. К задаче прогнозирования продаж методом регрессии мы перейдем немного позже, в разделе “Интерпретация коэффициентов и их стандартных ошибок”.)

Базовый набор данных на рис. 16.1 находится в столбцах `A`, `B` и `C`; результаты функции `ТЕНДЕНЦИЯ` выведены в столбец `E`. В столбце `E` мы видим прогнозы значения веса для заданных в строке значений возраста и роста, при этом ряд этих прогнозов имеет максимальную из возможных корреляцию с рядом значений фактического веса.

Если это все, что вас интересует в прогнозировании, можете остановиться на этом месте. И это можно вполне понять, так как вы уже получили прогноз, которого добивались. Однако это далеко не все — нужно взглянуть и на другую информацию в результатах регрессионного анализа. Из нее можно почерпнуть следующее.

- ✓ Имеет ли смысл выполнять прогноз веса по росту и возрасту.
- ✓ Следует ли в качестве предикатов использовать обе переменные (роста и возраста) или можно обойтись какой-либо одной из них.
- ✓ Насколько достоверным окажется уравнение регрессии, если использовать его для других множеств исходных данных.

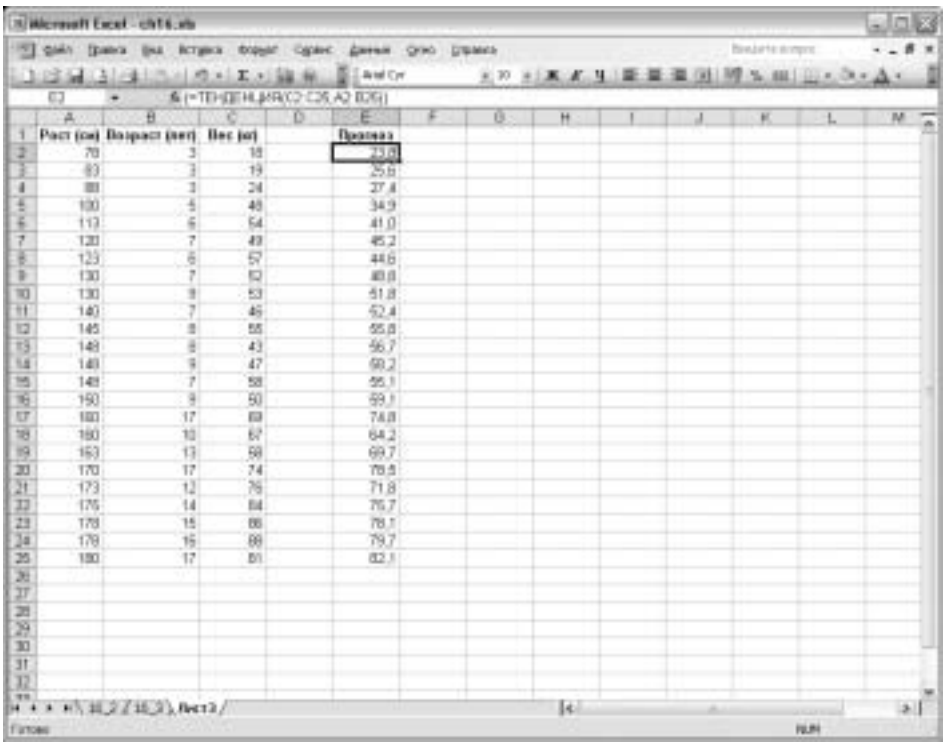


Рис. 16.1. Вы можете ввести функцию ТЕНДЕНЦИЯ как массив, нажав после ввода комбинацию клавиш <Ctrl+Shift+Enter>, как и в случае с функцией ЛИНЕЙН

Ответы на эти вопросы чрезвычайно важны, поскольку они помогут принять решение, использовать ли полученное уравнение регрессии, и если использовать, то как. Возможно, целесообразно выбрать в качестве предикатов какие-то другие переменные. Вы увидите, насколько устойчивым является полученное уравнение, и многое другое. Результаты более полного регрессионного анализа показаны на рис. 16.2.

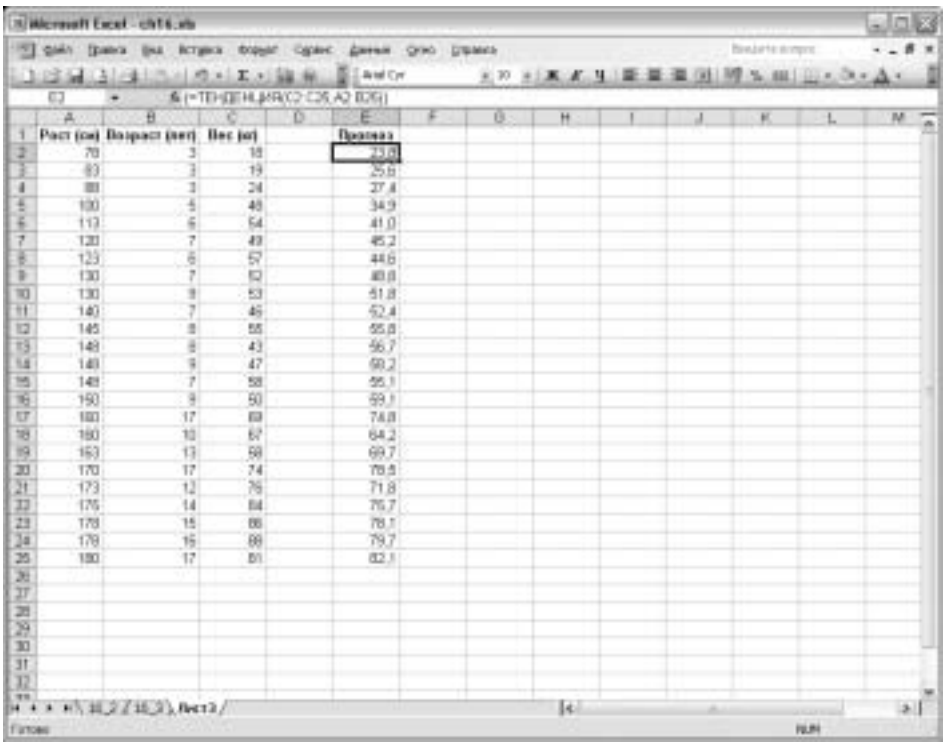
На этом рисунке мы видим информацию, полученную с помощью функции ЛИНЕЙН (LINEST), которая важна для интерпретации множественной регрессии. Ниже описаны действия, выполненные для получения результатов, показанных на рис. 16.2. Если имеющиеся у вас данные расположены в других столбцах или строках, просто откорректируйте адреса, указанные в пп. 2 и 3, соответствующим образом.

1. **Подсчитайте количество переменных предиката и добавьте к результату единицу**.

   На рис. 16.2 мы располагаем двумя переменными предиката — ростом и возрастом, значит, результатом будет число 3.

2. **Выделите диапазон ячеек (свободных или с уже не интересующими вас данными), содержащий пять строк и полученное на первом шаге число столбцов**.

   В нашем примере мы выделяем диапазон ячеек размерности пять строк на три столбца — E2:G6.



	A	B	C	D	E	F	G	H	I	J	K	L	M
1	Рост (см)	Возраст (лет)	Вес (кг)		Прогноз								
2	70	3	10		23,6								
3	83	3	19		25,6								
4	88	3	24		27,4								
5	100	5	48		34,9								
6	112	6	54		41,0								
7	120	7	49		45,2								
8	123	6	57		44,6								
9	130	7	62		48,8								
10	130	8	53		51,8								
11	140	7	46		52,4								
12	145	8	55		55,8								
13	148	8	43		56,7								
14	148	9	47		59,2								
15	148	7	58		56,1								
16	150	9	50		59,1								
17	160	17	69		74,8								
18	160	10	67		64,2								
19	163	13	58		69,7								
20	170	17	74		78,8								
21	173	12	76		71,8								
22	175	14	84		76,7								
23	178	15	86		78,1								
24	178	16	89		79,7								
25	180	17	81		82,1								
26													
27													
28													
29													
30													
31													
32													

Рис. 16.2. Фигурные скобки, в которые заключена формула, указывают на ее ввод в режиме массива

3. **В строке формулы введите** `=ЛИНЕЙН(C2:C25;A2:A25,,ИСТИНА)`**, но не нажимайте клавишу <Enter>**.
4. **Нажмите комбинацию клавиш <Ctrl+Shift+Enter>**.

   Этот метод позволяет ввести формулу в режиме массива. ЛИНЕЙН — это одна из функций Excel, для получения корректных результатов *требующая* своего ввода в режиме массива.

Если, как в этом случае, предполагается ввод формулы в режиме массива, начинайте операцию с выделения полного диапазона ячеек результатов. Программа Excel не занимается самостоятельно этим вопросом — она использует данные, предоставленные вами. В целом, в программе Excel существует несколько таких узких мест, способных поставить пользователя в затруднительное положение, и это одно из них.

Функция ЛИНЕЙН позволит вам увидеть следующее.

- ✓ В первой строке содержатся *коэффициенты* и постоянная уравнения регрессии (называемая *пересечением*). Эти числа вместе с фактическими данными используются для вычисления прогноза. На рис. 16.2 они находятся в ячейках `E2:G2`.
- ✓ Вторая строка содержит *стандартные ошибки* коэффициентов и пересечения. Эти числа помогут вам принять решение, каким переменным уделять внимание при создании прогноза. На рис. 16.2 эти значения содержатся в ячейках `E3:G3`.

- ✓ В следующих строках полезная информация находится в первых двух столбцах; третий столбец содержит только сообщения об ошибке (#Н/Д). На рис. 16.2 нас могут заинтересовать ячейки E2:F6.

Что представляет собой эта полезная информация? Разрешите предложить краткий обзор (именно *краткий*, поскольку в литературе по статистике каждому из этих значений обычно посвящают целую главу).

R-квадрат

В первом столбце третьей строки результатов функции ЛИНЕЙН содержится значение *R-квадрат*. Это квадрат коэффициента корреляции между фактическими данными и прогнозом. В примере на рис. 16.2 это квадрат корреляции между значениями в ячейках C2:C25 (фактические) и A2:B25 (прогноз). В данном случае значением "R-квадрат" является число 0,84.

Значение "R-квадрат" на самом деле является процентом вариаций — *рассеивания* — прогнозируемой переменной, который определяется (характеризуется) переменными предиката. В данном примере 84% фактических значений веса может быть определено с помощью комбинации роста и возраста. Исходя из этого б*о*льший вес ассоциируется с большими значениями роста и возраста, меньший вес — с меньшими.

Если это кажется вам похожим на определение корреляции, вы не ошиблись. Свое название значение "R-квадрат" получило по той причине, что оно является квадратом коэффициента корреляции (коэффициенту корреляции принято присваивать имя *r* (в простой корреляции) или *R* (в множественной)).

Чем выше значение "R-квадрат", тем лучше свою работу в качестве предсказателей выполняют переменные предикатов. Так, это значение является квадратом, оно по определению не может быть отрицательным, а так как сама корреляция по абсолютному значению не превосходит единицы, то областью значений "R-квадрат" можно считать диапазон от нуля до единицы.

Таким образом, получив результаты функции ЛИНЕЙН, в первую очередь, обратите внимание на ячейку в третьей строке и первом столбце. Чем ближе это число к единице, тем лучше прогноз; чем ближе к нулю — тем хуже.

Никто не совершенен…

И функция ЛИНЕЙН в том числе. В ячейке в третьей строке и втором столбце ее результатов содержится значение *стандартной ошибки оценки*, позволяющей оценить погрешность прогнозирования.

На рис. 16.2 стандартную ошибку оценки можно найти в ячейке F4, и равна она 8,25. Если отложить от любого прогноза два значения стандартной ошибки вверх и вниз, то эти числа ограничат диапазон, в который с 95% вероятностью попадают фактические значения наблюдений.

Для примера предположим, что нам требуется предсказать вес 13-летнего юноши ростом 150 см. Значения, полученные функцией ЛИНЕЙН на рис. 16.2, позволяют составить следующее уравнение:

```
-8,02+(1,58×Возраст)+(0,35×Рост)=Прогноз_веса
-8,02+(1,58×13)+(0,35×150)=65,0
```

В главе 12 уже отмечалось, что в результатах функции ЛИНЕЙН коэффициенты предикатов расположены в порядке, *обратном* тому, который присутствует в базовом наборе данных. Именно по этой причине коэффициент `1,58` использован для возраста, а `0,35` — для роста, в то время как на рабочем листе столбец роста предшествует столбцу возраста.

Итак, стандартная ошибка оценки в данном случае равна `8,25`. Следовательно, при составлении этого прогноза вы на 95% можете быть уверены, что фактический вес будет находиться в пределах от `48,5` до `81,5` килограммов:

```
65,0 – 2×8,25 = 48,5
65,0 + 2×8,25 = 81,5
```

Люди часто неверно интерпретируют фразу “на 95% уверены”. Это совсем не означает, что вероятность попадания фактического значения в диапазон между верхним и нижним пределами (т.е. между 48,5 и 81,5 килограммами) равна 95%. Вероятность может быть равна единице (если фактическое значение попало в диапазон) или нулю (если не попало). В то же время, если применить эту меру к тысячам людей, окажется, что 95% из них будут иметь вес, находящийся в указанных нами пределах.

Как вы можете догадаться, чем выше значение “R-квадрат”, тем меньше стандартная ошибка оценки (мера неточности прогноза).

Прочая статистика функции ЛИНЕЙН

В первых двух столбцах четвертой и пятой строк результатов функции ЛИНЕЙН содержатся понятные лишь немногим данные. Если вы не имеете специальной подготовки в области статистического анализа, смело можете их игнорировать. Эти данные являются строительными блоками дальнейшего анализа, позволяющего определить *статистическую значимость* полученных результатов.

К этой группе данных можно проявить разный интерес и предпринять соответствующие действия, в частности следующие.

- ✓ При **малом** интересе эти данные можно игнорировать и использовать для получения прогноза функцию ТЕНДЕНЦИЯ.
- ✓ При **среднем** интересе можно использовать функцию регрессии пакета анализа и получить из нее информацию об уравнении регрессии. Эта функция выполняет и более глубокий анализ. В главе 11 мы уже подробно останавливались на результатах этой функции.

- ✓ При **высоком** интересе, если вы считаете себя достаточно подготовленным теоретически, можете обратить внимание на следующие данные.
  - **F-статистика**. Это значение получается делением среднеквадратического значения регрессии на среднеквадратическое отклонение. Оно находится в первом столбце четвертой строки результатов функции ЛИНЕЙН. Если использовать функцию FРАСП (FDIST) программы Excel, задав в ней F-статистику, количество переменных предиката и степень свободы остатков (см. далее), то можно вычислить статистическую значимость регрессии. Это равнозначно тестированию, насколько сильно отличается R-квадрат от нуля.

- **Степень свободы**. Это значение находится во втором столбце четвертой строки результатов функции ЛИНЕЙН. Оно указывает, на что нужно разделить сумму квадратов остатков, чтобы получить среднеквадратическое отклонение. Это значение является третьим аргументом функции FРАСП.
- **Сумма квадратов регрессии**. Это значение находится в первом столбце пятой строки результатов функции ЛИНЕЙН. Делением этого числа на количество переменных предикатов можно получить среднеквадратическое значение регрессии.
- **Сумма квадратов остатков**. Это значение находится во втором столбце пятой строки результатов функции ЛИНЕЙН. Если разделить сумму квадратов регрессии на это число, можно получить значение F-статистики.

Обдуманный подход к множественной регрессии

В предыдущем разделе уже говорилось, что функция регрессии пакета анализа всего лишь наполовину выполняет задачу прогнозирования. Она позволяет получить результаты, предлагаемые функцией ЛИНЕЙН, как для создания прогноза, так и для диагностики того, насколько хорошо уравнение регрессии. Так что, если вы решили использовать в прогнозировании функцию регрессии пакета анализа, вы не можете полностью игнорировать ее результаты.

В то же время вы можете решить не проводить всю эту диагностику собственноручно, основываясь на результатах функции ЛИНЕЙН. Конечно, это можно сделать и вручную, однако такое решение можно отнести к разряду приемлемых только для пуристов.

В главе 11 был проведен обзор различных элементов результата функции регрессии пакета анализа, которые критичны для принятия решения относительно представления результатов прогноза для широкой аудитории. Позволю еще раз повторить его, сократив до минимума описания.

- ✓ **Множественный R**. Это квадратный корень из значения “R-квадрат”. Для оценки точности уравнения регрессии можно использовать любую из этих величин. Если вам более удобно работать в терминах коэффициентов корреляции, обратите внимание на значение множественного R.
- ✓ **Пересечение и коэффициенты**. Эти числа применяются к значениям предикатов для получения наиболее достоверного из возможных прогноза.
- ✓ **Уровни достоверности**. Эти окрестности точки пересечения и коэффициентов помогают оценить величину разброса фактических значений вокруг значений прогнозов в других аналогичных наборах данных.
- ✓ **График остатков**. Посмотрите на этот график и оцените, существуют ли общие закономерности в ошибках прогнозирования.
- ✓ **График подбора**. Это график прогнозируемых и фактических значений относительно значений предиката. В множественной регрессии этот график не так полезен, поскольку позволяет использовать одновременно только один предикат.

Неправдоподобное сжатие R-квадрата

На рис. 16.3 показан пример результатов функции регрессии пакета анализа на том же наборе исходных данных, на котором были получены результаты функции ЛИНЕЙН на рис. 16.2.

Значение *Нормированный R-квадрат* находится в ячейке `F6` на рис. 16.3. В задачах прогнозирования продаж ему уделяют повышенное внимание.

Предположим, что мы собираемся получить данные о росте и возрасте другой группы молодых людей и на их основе составить прогноз их веса с помощью уже имеющихся коэффициентов уравнения регрессии, находящихся в ячейках `F17:F19` на рис. 16.3.

В этом случае нужно узнать корреляцию полученных новых прогнозов с новыми фактическими данными (то, что на рис. 16.3 мы узнали из значения множественного R для исходного набора данных).

Новое значение множественного R, вероятнее всего, будет меньше исходного — оно *сжимается*. Причины этого явления несколько туманны, однако они имеют отношение к обобщению исходной выборки данных. Для чего нам это нужно? На то существует ряд веских причин.

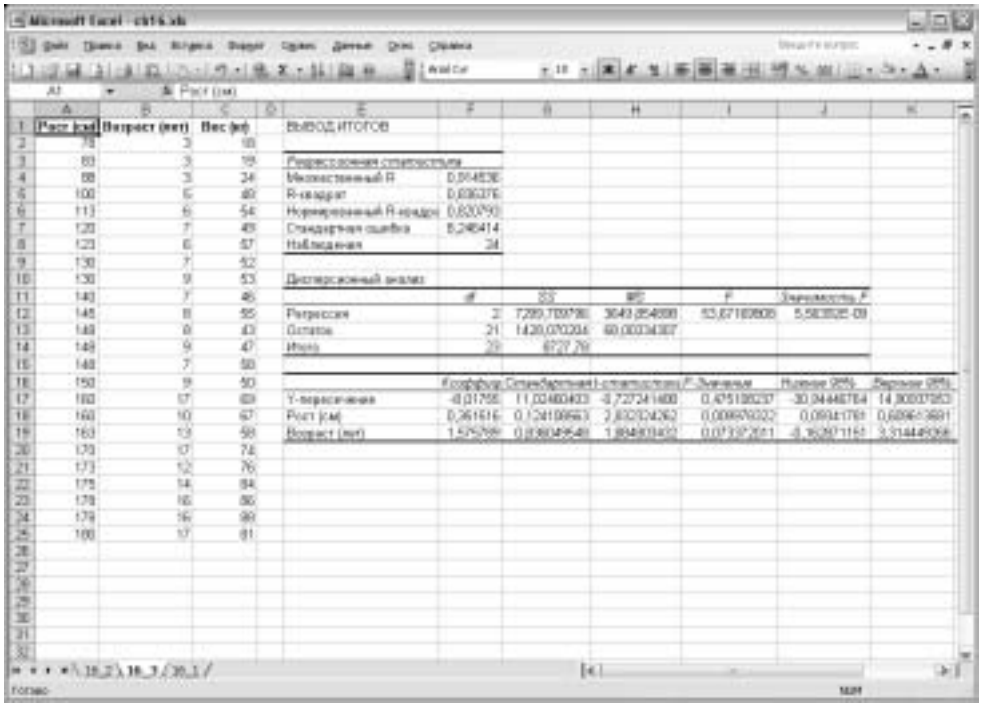


Рис. 16.3. Функция регрессии пакета анализа вооружает вас большим объемом информации, чем ЛИНЕЙН, однако и она имеет свои недостатки

В прогнозировании продаж мы имеем дело с постоянно поступающими новыми данными. Вскоре после того, как был создан прогноз на июль или второй квартал, начинают поступать новые фактические данные и приходит время заняться прогнозом на следующий период (август или третий квартал). Естественно, при этом вы захотите сравнить эти новые фактические данные с ранее созданным прогнозом.

Если уделить внимание значению нормированного R-квадрата, вы получите нечто вроде пессимистической оценки точности регрессионного прогноза на следующий период. Следует учесть, что в этом случае данные не обновляются полностью — к ним добавляется ряд новых значений. Если значение нормированного R-квадрата все еще при-

емлемо для вас, значит, вы можете с определенной долей уверенности ожидать точности от нового прогноза.

Еще одной важной причиной повышенного внимания к значению нормированного R-квадрата является то, что оно чувствительно к взаимосвязи между количеством наблюдений в базовом наборе данных и количеством предикатов в уравнении регрессии. Вот его формула:

```
Нормированный R-квадрат = 1 – [(1–R-квадрат)×([N–1]÷[N–K–1])]
```

Здесь *N* — количество наблюдений, а *K* — количество переменных предикатов в анализе.

Чтобы увидеть результаты работы этой формулы на множестве данных о продажах, посмотрите на рис. 16.4. На этом рисунке функция регрессии пакета анализа была выполнена на множестве данных об объеме продаж, затратах на продвижение товара, размере штата продавцов и скидок, предлагаемых покупателям (эти скидки были нацелены на уменьшение складских запасов). Результаты функции регрессии показаны в ячейках `F1:L20`.

Применив формулу нормированного R-квадрата к ячейке `G23`, мы получили число `0,56`. Это же значение мы видим в результатах функции регрессии (ячейка `G6`). Добавление дополнительных переменных в уравнение вызывает быстрое падение нормированного R-квадрата — до значения `0,46` для шести переменных (ячейка `G26`).

Теперь принимаем во внимание то, что в анализе "все прочие условия остаются равными". К примеру, если некоторая одна добавленная в уравнение регрессии переменная имеет корреляцию с прогнозируемой, равную единице, то и значение "R-квадрат" "подпрыгнет" до единицы, сделав лжецами нормированный R-квадрат и меня. Однако сначала найдите такую переменную, которая способна идеально прогнозировать продажи, и познакомьте с ней меня.

Подсчет по головам

В результатах функции регрессии пакета анализа мы видим также и подсчет количества наблюдений (в контексте прогнозирования продаж это количество записей в базовом наборе данных), которые были приняты в расчет в анализе. На рис. 16.4 это значение вы найдете в ячейке `G8`. Оно может использоваться для вычисления значения нормированного R-квадрата — обратите внимание на формулу в ячейке `G23` на рис. 16.4.

Степени свободы

Для нас представляют интерес два значения: степень свободы регрессии и степень свободы остатков. Чтобы понять, *почему* степени свободы играют такую важную роль, нужно изучить массу материала и изрядно поломать голову. Поэтому сейчас мы ограничимся только констатацией факта, что для получения среднеквадратического значения нужно разделить сумму квадратов (ячейки `H12` и `H13`) на соответствующую степень свободы.

Среднеквадратическое значение

Это синоним термина *вариация*. Для получения F-статистики нужно разделить среднеквадратическое значение регрессии на среднеквадратическое значение остатков.

F-статистика

F-статистика является именно тем значением, на которое смотрят при определении значимости регрессии; говоря более точно, при определении, насколько сильно значение "R-квадрат" отличается от нуля. Этот вопрос более подробно был освещен в разделе "Прочая статистика функции `ЛИНЕЙН`".

Рис. 16.4. Таблица результатов содержит статические значения, а не формулы, поэтому при изменении исходных данных функция регрессии должна быть запущена повторно

Значимость F

Говоря ранее о функции ЛИНЕЙН, мы упоминали функцию FРАСП и то, как ее можно использовать наряду со значениями статистики F и степеней свободы для определения того, насколько вы можете быть уверенными, что истинный R-квадрат больше нуля. Функция регрессии пакета анализа выполняет эту работу за вас. На рис. 16.4 уровень значимости вы можете найти в ячейке K12. Чем меньше уровень значимости, тем сильнее вы можете быть уверены, что получили статистически значимую регрессию.

На первый взгляд, это может выглядеть очень важным, и, возможно, это так и есть. Однако на самом деле это значит, что истинный R-квадрат, вычисленный на всем множестве генеральной совокупности (т.е. на всем *потенциально* доступном множестве наблюдений), отличен от нуля.

Вся эта эзотерическая статистика — степени свободы, среднеквадратические значения, F-статистика и прочие — вычисляется в функции регрессии пакета анализа по соглашению. Все эти значения были введены в теории вариационного анализа сэра Рональда Фишера, и поэтому отображаются в отдельной таблице с названием *ANOVA* (*Analysis Of Variations*) (см. ячейку F10 на рис. 16.4). Естественно, неплохо знать все эти числа, в прогнозировании продаж можно извлечь пользу только из значимости F (для оценки достоверности уравнения регрессии).

Интерпретация коэффициентов и их стандартных ошибок

В заключительном разделе, посвященном результатам функции регрессии пакета анализа, мы рассмотрим значения коэффициентов и пересечения, используемые для создания прогноза. На рис. 16.4 пересечение содержится в ячейке `G17`, а коэффициенты — в ячейках `G18:G20`.

Несмотря на свое нахождение в столбце `Коэффициенты`, Y-пересечение не является коэффициентом. Это число, которое нужно добавить к уравнению регрессии в качестве коррекции, а коэффициенты — это числа, на которые нужно умножить в уравнении переменные предикатов.

Стандартные ошибки, ассоциированные с коэффициентами и Y-пересечением, помогают оценить охват уравнением области значений. Эти стандартные ошибки используются подобно стандартным ошибкам оценки, о которых говорилось в разделе "Никто не совершенен...". Добавление двух стандартных ошибок к коэффициенту и вычитание двух стандартных ошибок из него позволяют увидеть, захватывает ли полученный диапазон нуль. Величина в две стандартные ошибки оценивает разброс значений коэффициента. При желании вы можете построить область, ограниченную величиной в три стандартные ошибки. На рис. 16.5 показано, как этот метод работает с данными о продажах, показанными на рис. 16.4.

Все столбцы в диапазоне `C2:I6` свидетельствуют об одном — только переменная скидки достоверно связана с объемом продаж, по крайней мере в представленном исходном наборе данных.

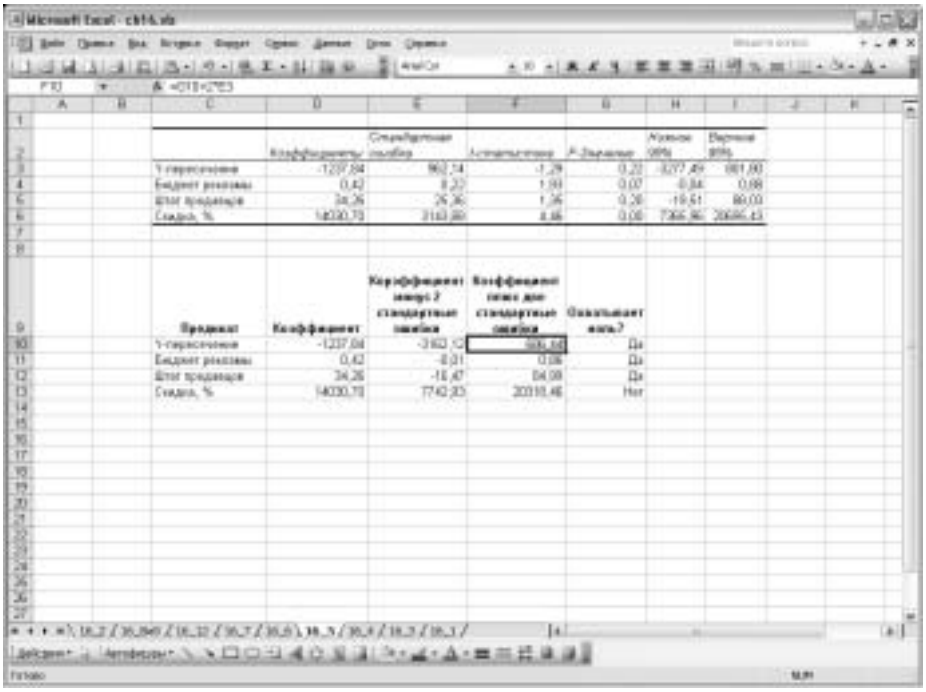


Рис. 16.5. Если диапазон значений коэффициента захватывает нуль, нужно учесть вероятность, что он действительно равен нулю

В частности, если вычесть и добавить стандартные ошибки к коэффициентам и Y-пересечению, все результирующие диапазоны будут включать в себя нуль, кроме коэффициента переменной скидки. Это значит, что нельзя гарантировать то, что на полном наборе данных (генеральной совокупности) эти коэффициенты не будут нулевыми. А если эти коэффициенты станут нулевыми, то уравнение

```
Объем_продаж=-1227,58+(0,42×Реклама)+(34,74×Активные_продажи)+
(13983,36×Скидка)
```

превратится в

```
Объем_продаж=-0+(0×Реклама)+(0×Активные_продажи)+(13983,36×Скидка)
```

Это значит, что переменные затрат на продвижение товара и штат продавцов, а также Y-пересечение следует выбросить из уравнения регрессии, так как нет никаких оснований верить, что истинные коэффициенты и истинное Y-пересечение ненулевые.

В главе 4 говорилось, что основной чертой регрессионного подхода в прогнозировании является скупость: чем меньше переменных предикатов в уравнении регрессии, тем лучше. Исходя из этого принципа, если избавиться от переменных затрат на рекламу и штата продавцов (и в результате уравнение регрессии окажется хорошим), то это к лучшему. По крайней мере, вам не придется ежемесячно или ежеквартально собирать отсеченную информацию. Возвращаясь к нашему разговору о нормированном R-квадрате и его сжатии, следует сказать, что уменьшение количества предикатов полезно и в этом отношении.

На рис. 16.6 показан другой взгляд на эти вещи. В одном из примечаний этого раздела говорилось, как можно использовать информацию в результатах функции регрессии пакета анализа для определения достоверности коэффициентов. Только что мы рассмотрели диапазоны коэффициентов, основанные на стандартных ошибках. Сейчас мы остановимся на t-статистике и доверительных интервалах. Они могут привести нас к таким же выводам, а если нет, следует пристальнее всмотреться, о чем эти значения нам говорят.

T-статистика поможет определить статистическую значимость отличия числа от нуля. В данном случае она вычисляется делением числа (пересечения или коэффициента) на его стандартную ошибку. На рис. 16.6 эта формула введена в ячейки `F10:F13`. И мы видим результаты, не отличающиеся от значений t-статистики, в результатах функции регрессии (ячейки `E3:E6` на рис. 16.6).

Функция регрессии вычисляет также статистическую значимость t-статистики; эти результаты показаны в ячейках `F3:F6`. Эти результаты можно интерпретировать таким же образом, как и величину `Значимость F` (см. одноименный раздел ранее в этой главе).

В нашем примере функция регрессии вернула p-значение, существенно превышающее величину 0,05 для переменных затрат на рекламу и штата продавцов (именно величина 0,05 считается критерием статистической значимости). И только одна переменная пересекла этот рубеж — величина скидки. Это говорит нам о том, что ее корреляция с объемом продаж *существенно* выше нуля.

Чем *меньше* P-значение (в таблице ANOVA оно называется значимостью F), тем *более* значимой является соответствующая переменная.

На рис. 16.6 в ячейках `H10:H13` P-значения были вычислены также с помощью функции `СТЬЮДРАСП` (`TDIST`). Обратите внимание, что они практически идентичны тем, которые вернула функция регрессии пакета анализа в своих результатах.

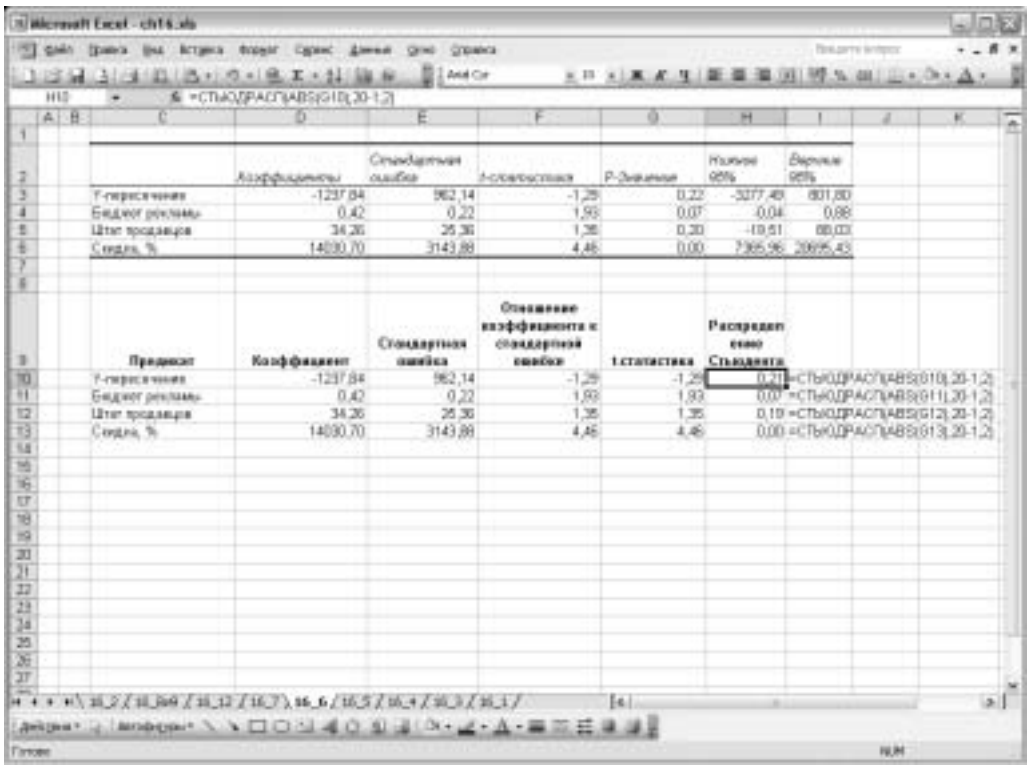


Рис. 16.6. Сравнение значений в столбце "P-Значение" со значениями в столбце "Распределние Стьюдента"

Функция СТЬЮДРАСП предполагает наличие неотрицательного аргумента. Если вы собираетесь использовать функцию СТЬЮДРАСП, передавайте ей абсолютное значение аргумента, применяя при этом функцию ABS. Абсолютные значения всегда положительны, и это гарантированно позволит вам избежать ошибок. Если бы в ячейках H10:H13 на рис. 16.6 не использовалась функция ABS, то результатом функции СТЬЮДРАСП была бы ошибка, если бы первый ее аргумент был отрицательным.

Итак, анализ с помощью t-статистики привел нас к тому же выводу, что и проверка попадания нуля в область значений коэффициентов. Мы увидели, что для прогноза объема продаж вполне достаточно переменной скидки, веских же аргументов в пользу использования остальных переменных мы так и не нашли.

В заключение мы можем исследовать значение доверительного интервала, возвращаемое функцией регрессии пакета анализа (рис. 16.7).

Как и следовало ожидать исходя из предыдущих анализов, доверительный интервал предикатов бюджета рекламы и штата продавцов включает в себя нуль. В каждом из этих случаев значение верхних 95% больше нуля, а нижних 95% — меньше. Следовательно, нельзя исключить вероятность того, что истинные коэффициенты и Y-пересечение всей генеральной совокупности являются нулевыми. Это нас еще раз убеждает в том, что наилучший прогноз можно получить, используя только предикат скидки.

Разница между верхними и нижними 95% и интервалом, полученным с помощью откладывания двух стандартных ошибок, заключается в том, что первое значение было основано на t-значении 2,1. Причины этого объяснить довольно сложно, однако нужно принять во внимание тот факт, что размер исследуемой выборки исходных данных весьма ограничен. Это было бы не так, если бы мы имели в своем распоряжении всю генеральную совокупность значений переменных предикатов.

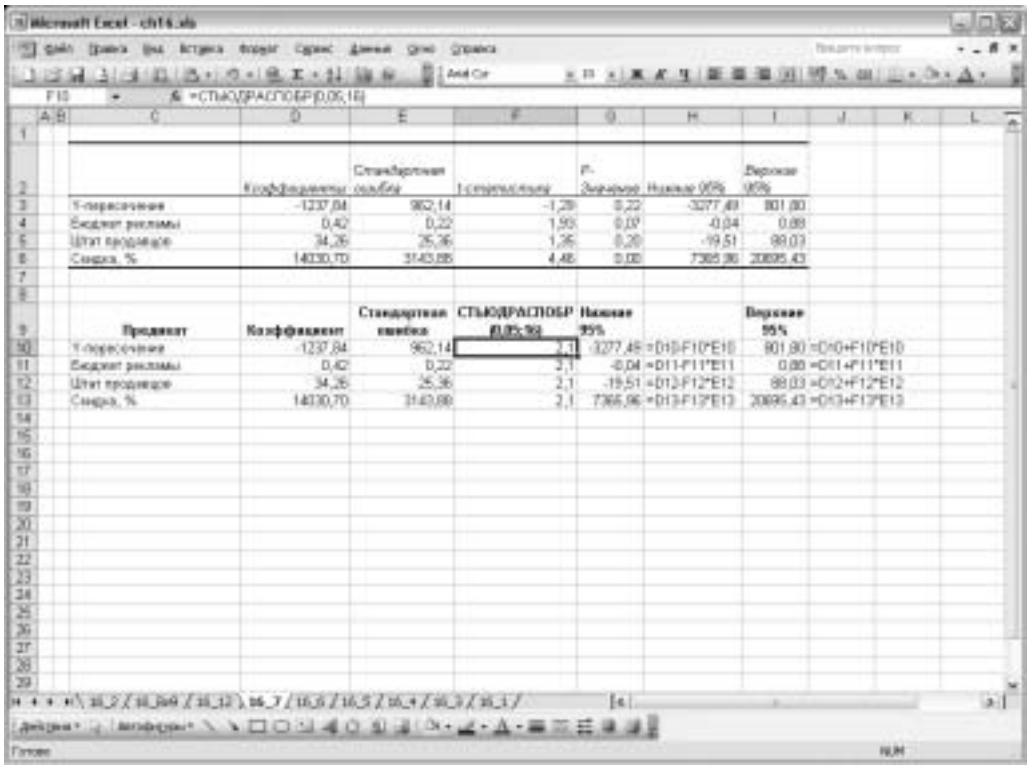


Рис. 16.7. Использование статистики Стьюдента для создания 95% интервалов

Значения верхних и нижних 95% можно вычислить с помощью функции СТЬЮДРАСПОБР (TINV), что и показано на рис. 16.7 в ячейках `F10:F13`. В качестве аргументов мы передаем в эту функцию `0,05` (т.е. 1–0,95) и число степеней свободы и в результате получаем значение `2,1`. Умножая 2,1 на величину стандартной ошибки и прибавляя результат к коэффициентам и Y-пересечению, мы получаем верхние 95%. Для получения нижних 95% результат операции умножения нужно отнимать.

В рассмотренном примере все три теста привели нас к одному и тому же заключению: в качестве предиката нужно использовать только одну переменную — скидку, и при этих условиях нужно пересчитать Y-пересечение. Существует совсем немного ситуаций, когда описанные методы анализа не согласуются друг с другом. Если вы столкнулись с такой ситуацией, возможно, нужно увеличить базовый набор данных, либо расширив его назад, в прошлое, либо дождавшись очередных новых фактических данных. Большие наборы исходных данных, как правило, решают проблему неопределенности.

Также примите к сведению, что формулы, показанные на рис. 16.5–16.7, будут автоматически пересчитываться при изменении исходных данных. Именно по этой причине я

отдаю предпочтение формулам, введенным вручную, а не функции регрессии пакета анализа, возвращающей статические значения.

Помещение линии тренда на график

Помещение линии тренда на диаграмму позволяет визуально оценить полученный прогноз и сравнить его с фактическими данными. Это и является главной задачей линии тренда, и в этом разделе будет показано, как ее создать.

Никогда не было хорошей идеей смотреть только на голые цифры, сгенерированные компьютером. Чтобы понять, имеет ли смысл полученный прогноз, нужно воспользоваться какими-либо статистическими методами. В предыдущем разделе уже было показано, как в программе Excel получить статистические данные, позволяющие сделать взвешенные заключения.

Однако голые цифры не смогут рассказать вам все. На рис. 16.8 в качестве примера показана ситуация, в которой цифры могут сбить с толку, но одного взгляда на график достаточно, чтобы ситуация прояснилась.

На этом рисунке явно прослеживается взаимосвязь количества торговых представителей и объема продаж. До определенного предела с увеличением количества представителей повышается и доход. Однако как только число торговых представителей превышает 16, доходность начинает резко падать. И этому можно найти ряд объяснений.

- ✓ Регионы продаж могут быть определены как по территориальному, так и по национальному признакам. Вполне возможна ситуация, когда в одном географическом регионе сталкиваются интересы двух разных торговых представителей, и в погоне за комиссионными они начинают конкурировать друг с другом, даже несмотря на то, что работают на одну компанию.
- ✓ Для охвата любого региона продаж вполне достаточно определенного числа торговых представителей.
- ✓ В игру вступила некоторая другая переменная. К примеру, руководство могло прийти к решению увеличить число торговых представителей, чтобы усилить свое присутствие на рынке, хотя стоило бы поработать над потребительскими качествами самого товара.

Коэффициент корреляции, о котором мы говорили в главе 14, является примером *линейной статистики*. Другими словами, он предполагает наличие линейной связи между двумя переменными (как в случае зависимости веса от роста). Эта связь не обязана быть идеальной, однако в общем случае увеличение значения одной переменной должно приводить к увеличению и второй.

Зависимость переменных может быть и обратной, как в случае статистики дорожных аварий и возраста водителей. Чем старше водитель, тем больше его опыт и тем реже он попадает в аварии (по крайней мере, до 30-летнего возраста). В результате корреляция этих переменных будет отрицательной и в то же время сильной.

На рис. 16.8 в ячейке `G1` показано значение корреляции между количеством торговых представителей и объемом продаж. Как мы видим, она близка к нулю. Однако нелинейная взаимосвязь переменных, измеренная статистикой с названием *эта-квадрат* или *от-*

ношение корреляции, в ячейке G2 довольно сильна, что мы и видим на графике. Однако в данном случае взаимосвязь уже не является линейной.

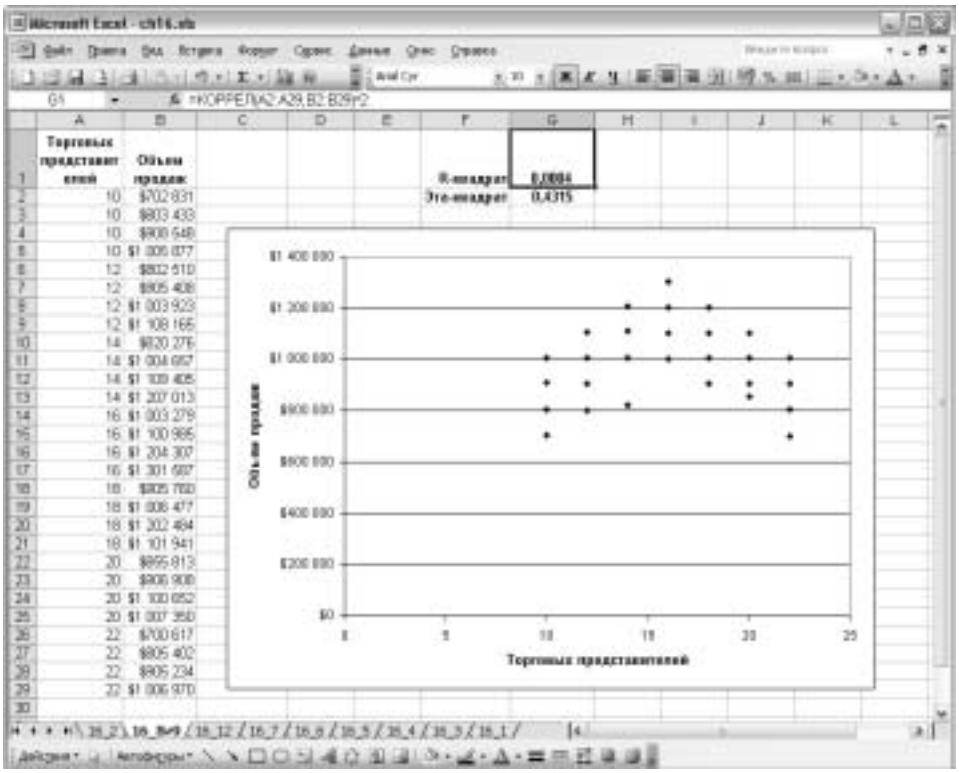


Рис. 16.8. Эта-квадрат является обобщенной версией R-квадрата, возвращающей такое же значение для линейной зависимости, однако более точное в случае нелинейной

В ячейке G1 на рис. 16.8 введена формула вычисления квадрата коэффициента корреляции. Если значение R-квадрата вам приходится вычислять часто, может оказаться более удобным использовать функцию КВПИРСОН (RSQ). В примере на рис. 16.8 такая формула имела бы следующий вид:
=КВПИРСОН(A2:A29;B2:B29).

В то же время вместо игр с цифрами можно просто посмотреть на график. Линия тренда подскажет вам, насколько в данной ситуации уместен линейный анализ (рис. 16.9).

Для помещения на график линии тренда выполните следующие действия.

1. **Щелкните на диаграмме, сделав ее активной**.

   Обратите внимание, что при этом на линейке меню программы Excel добавляется новый пункт — Диаграмма (Chart).

2. **Выберите в меню пункт** Диаграмма⇨Показать линию тренда **(**Chart⇨Add Trendline**)**.

   Откроется диалоговое окно, показанное на рис. 16.10.

3. **Выберите тип** Линейная **(**Linear**) и щелкните на кнопке** ОК.

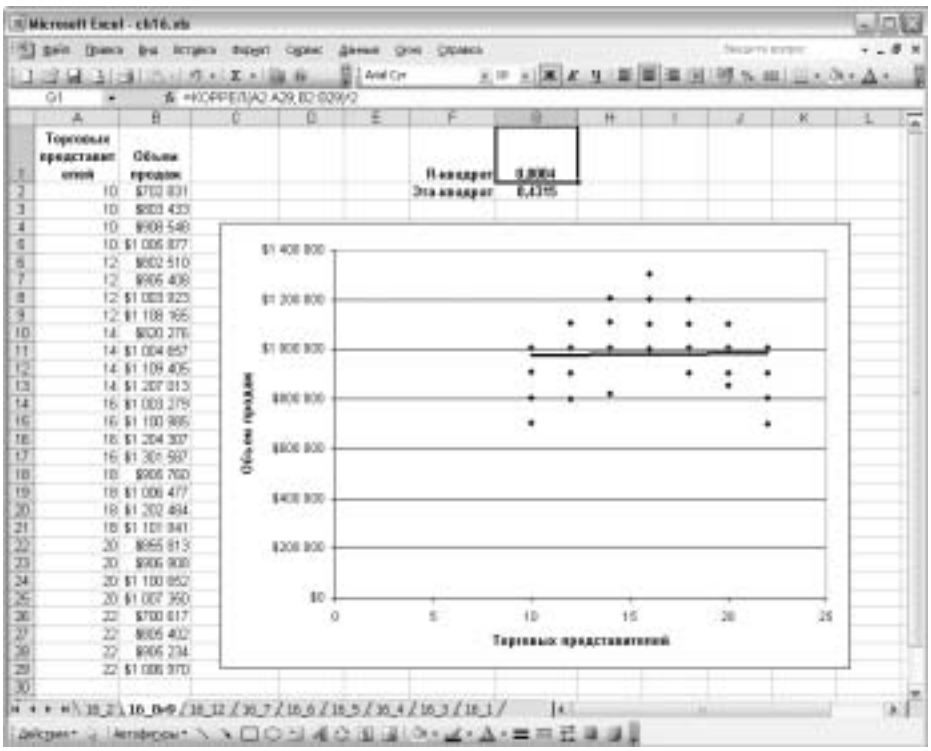


Рис. 16.9. Данный пример экстремален; даже не смотря на линию тренда, можно сказать, что зависимость переменных нелинейная

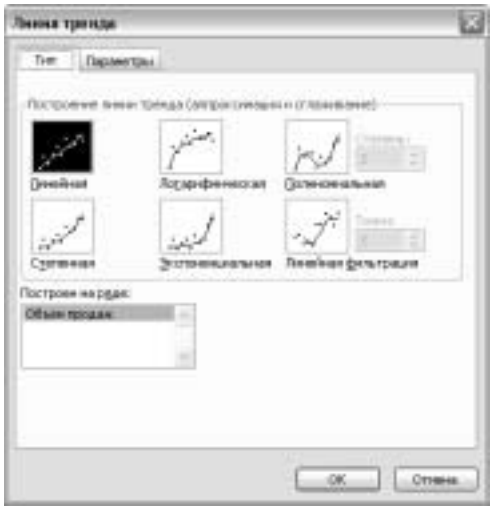


Рис. 16.10. Вы имеете возможность смоделировать базовый набор данных с помощью нелинейных линий тренда. Обычно для этого используют специальный статистический тест, называемый "качеством подобия"

Обратите внимание, что на рис. 16.9 линия тренда практически горизонтальная.

Если на диаграмме содержится несколько графиков разных рядов, перед созданием линии тренда выделите нужный график.

Когда между переменными существует сильная взаимосвязь, возможны две ситуации: наклон линии тренда близок к 45 градусам и точки на графике достаточно близки к линии тренда. Причиной этому является сам математический аппарат, лежащий в основе вычисления корреляции.

Если при создании линии тренда вас интересует и дополнительная информация, вы можете не останавливаться на п. 3 описанных выше действий, а продолжить их следующим образом.

1. **Щелкните на линии тренда, чтобы сделать ее активной**.
2. **Выберите в меню пункт** Формат⇨Выделенная линия тренда **(**Format⇨Selected Trendline**).**

   Откроется диалоговое окно Формат линии тренда.
3. **Перейдите во вкладку** Параметры **(**Options**), показанной на рис. 16.11**.

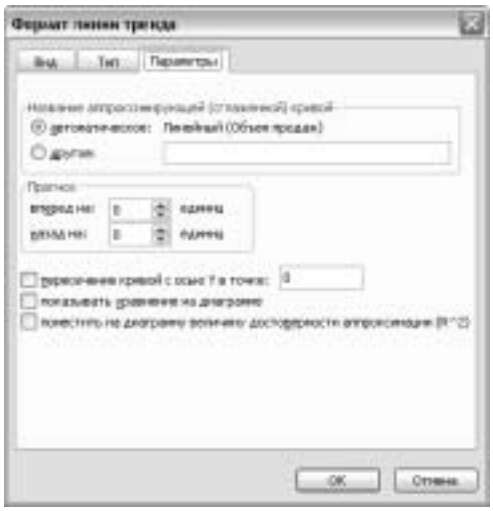


Рис. 16.11. Если на графике отображено несколько рядов данных, вернитесь ко вкладке Тип *и выберите нужный*

На этой вкладке доступно несколько параметров.

- ✓ При желании вы можете присвоить линии тренда более описательное имя, чем предложено программой по умолчанию. Это будет особенно полезно, если на диаграмме отображается легенда. Установите переключатель в положение Другое (Custom) и введите имя в текстовом поле.
- ✓ Если был выбран тип линии тренда, отличный от скользящего среднего, вы можете расширить линию тренда вперед, в будущее, или назад, в прошлое. С помощью

стрелочек в разделе Прогноз (Forecast) выберите число периодов удлинения линии тренда в каждом из направлений.

- ✓ При желании вы можете установить точку пересечения линии тренда с осью Y в нужное значение.
- ✓ Существует возможность отобразить на диаграмме само уравнение линии тренда; для этого нужно установить соответствующий флажок.
- ✓ На диаграмме можно отобразить также и значение R-квадрата, для чего, опять-таки, нужно установить соответствующий флажок. (В данном случае термин "R-квадрат" применяется в том же смысле, что и в функции регрессии пакета анализа, — величина достоверности аппроксимации прогноза относительно переменной предиката.)

Позволю себе дать некоторые комментарии относительно перечисленных выше параметров.

- ✓ Установка точки пересечения с осью Y вручную не рекомендуется. Устанавливая точку пересечения в нуль, можно увеличить значение "R-квадрат", однако аналитики, которые этим занимались, искали огрехи в математическом аппарате метода регрессии. Если исследуемый набор данных действительно имеет нулевое Y-пересечение, в результате анализа будет получено или точное нулевое значение, или значение, близкое к нему.
- ✓ Уравнение линии тренда и значение "R-квадрат" можно перетащить мышью подальше от графика, чтобы они не перекрывали другие элементы диаграммы.
- ✓ Как в уравнении, так и в значении "R-квадрат" можно изменить количество десятичных знаков и размер шрифта. Для этого нужно щелкнуть на соответствующем элементе и выбрать в меню пункт Формат⇨Выделенные подписи данных (Format⇨Selected Data Labels). Во вкладке Число можно отрегулировать формат отображения данных, в том числе количество десятичных знаков, а во вкладке Шрифт — начертание и размер шрифта. Также вы можете воспользоваться кнопками Увеличить разрядность и Уменьшить разрядность панели инструментов форматирования программы Excel.

Некоторые люди пытаются создавать прогнозы на основе уравнения линии тренда. Такое действие ошибочно. В значениях коэффициентов и Y-пересечения не отображается достаточное количество десятичных знаков, к тому же при вводе этих чисел в ячейки рабочего листа несложно допустить опечатки. Гораздо проще и быстрее использовать функцию ЛИНЕЙН и получить коэффициенты уже в ячейках рабочего листа или функцию ТЕНДЕНЦИЯ и сразу получить значения прогноза. Эти функции были описаны в главе 12.

Оценка прогноза, полученного методом регрессии

Когда для получения прогноза используется метод регрессии, следует обратить внимание и на проблемы, еще не упомянутые в материале этой главы, в частности на независимость ошибок. Две самые серьезные проблемы также могут возникнуть при наличии в исходных данных авторегрессии или согласованных трендов.

Использование авторегрессии

Тема авторегрессии уже поднималась в нескольких главах этой книги. Если говорить коротко, авторегрессия предполагает использование одного подмножества значений базового набора данных для предсказания другого подмножества того же ряда. Авторегрессию проще увидеть на диаграмме, чем читать о ней, поэтому взгляните на рис. 16.12.

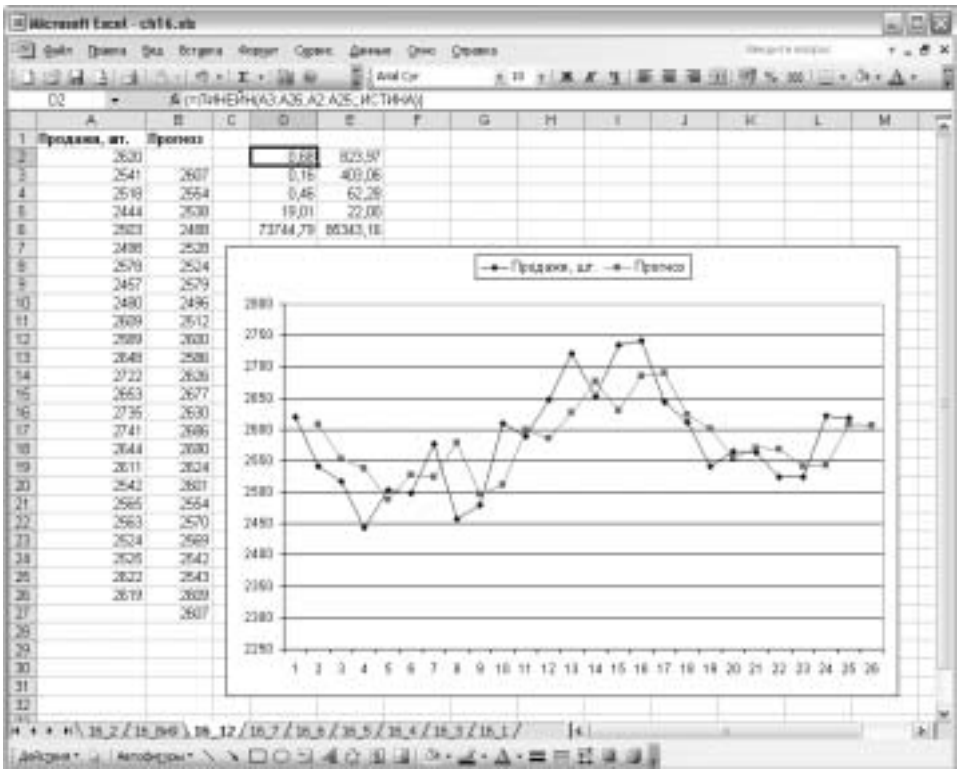


Рис. 16.12. Обратите внимание, что первый диапазон в аргументах функции ЛИНЕЙН описывает прогнозируемую переменную, что соответствует ее синтаксису

Функция ЛИНЕЙН на рис. 16.12 (ячейки `D2:E6`) использует значения в диапазоне `A2:A25` в качестве предиката, а значения со смещением на одну строку вниз (ячейки `A3:A26`) — в качестве прогнозируемой переменной. В результате мы как бы предсказываем будущие значения по прошлым значениям одного и того же ряда данных. Мы как бы указываем программе создать прогноз второго значения по первому, предполагая, что известна взаимосвязь значений в диапазонах `A2:A25` и `A3:A26`.

Авторегрессия похожа на экспоненциальное сглаживание, по крайней мере, в одном: значения предыдущих периодов помогают прогнозировать значения следующих.

Однако между этими методами есть и существенные различия. В частности, в методе экспоненциального сглаживания для создания прогноза на следующий период всегда используется предыдущее фактическое значение. В авторегрессии же для прогноза может использоваться не непосредственно предыдущее фактическое значение, а отстоящее от прогнозируемого на два или даже три периода времени.

С помощью результатов функции ЛИНЕЙН на рис. 16.12 мы можем создать прогноз, отображенный в столбце B. В частности, значение в ячейке B3 получено по формуле

```
=$E$2+$D$2*A2
```

Мы берем значение Y-пересечения и прибавляем его к произведению коэффициента регрессии и фактического значения в прошлом периоде. Далее мы копируем эту формулу и вставляем ее во все ячейки столбца B вплоть до B27.

Остались нерассмотренными два вопроса. Один из них связан с тем, что метод регрессии хорошо работает только со стационарными базовыми наборами данных; наличие тренда может спровоцировать неверные результаты. Второй вопрос связан с выбором интервала, на котором должно находиться значение предиката. На сколько периодов следует отступить? На один, два, а может, три или более? На рис. 16.12 был использован отступ в один период. Однако, чтобы обобщить ответ на эти вопросы, следует посмотреть на несколько графиков.

Зайдите на сайт www.dummies.com/go/excelsffd и загрузите рабочую книгу с названием Corellogram.xls. В ней содержится программа на языке VBA, анализирующая базовый набор данных и дающая ответ на вопрос относительно его стационарности. Также эта программа позволяет выбрать оптимальный интервал отступа значений предиката от прогноза (для использования в функции ЛИНЕЙН или ТЕНДЕНЦИЯ).

Когда вы откроете рабочую книгу Correlogram.xls, в меню Данные будет добавлен дополнительный пункт с названием Correlograms. Поместите базовый набор данных на первый лист рабочей книги (с названием Sheet1) и выберите в меню пункт Данные⇨Correlograms. Откроется диалоговое окно, в котором нужно задать диапазон ячеек, в котором расположен базовый набор данных. После щелчка на кнопке OK откроется новая рабочая книга с двумя диаграммами: ACF и PACF. Первая из них соответствует функции автокорреляции, вторая — функции частичной автокорреляции.

Пусть вас не смущают эти термины — вам они не понадобятся. Все, на что следует обратить внимание, — это сами графики На рис. 16.13 показана диаграмма ACF для множества данных, показанных на рис. 16.12.

На графике ACF показаны автокорреляции между двумя последовательностями наблюдений одного и того же базового набора данных, отстоящими друг от друга на один, два и более периодов времени. В стационарных наборах данных величина автокорреляции должна постепенно убывать до нуля и опускаться в область отрицательных чисел, что мы и видим на рис. 16.13.

Линии на рис. 16.13 и 16.14 (кривые в первом случае и прямые во втором) отображают предел статистической значимости. Все столбцы на графиках ACF и PACF, выходящие за пределы этого лимита, можно считать статистически значимыми, т.е. соответствующими достоверной методике.

Если полученный график ACF не соответствует показанной на рис. 16.13 модели, в которой автокорреляция стремительно падает, значит, мы имеем дело не со стационарным набором данных. В этом случае имеет смысл вычислить первые, а если этого окажется мало, и вторые конечные разности (как правило, первых оказывается достаточно; более подробно этот вопрос был освещен в главе 14).

На рис. 16.14 показан график PACF. Данный пример хорошо демонстрирует, насколько далеко можно отступить назад в авторегрессионном анализе. В данном случае первый столбик выходит за пределы лимита значимости (прямые линии на уровнях +0,4 и –0,4), а все остальные — нет. Это значит, что для создания прогноза следует использовать фактические данные, отстоящие от него на один период времени. То есть на рис. 16.12 нужно использовать функцию ЛИНЕЙН с диапазоном прогноза A3:A26 и диапазоном предиката A2:A25.

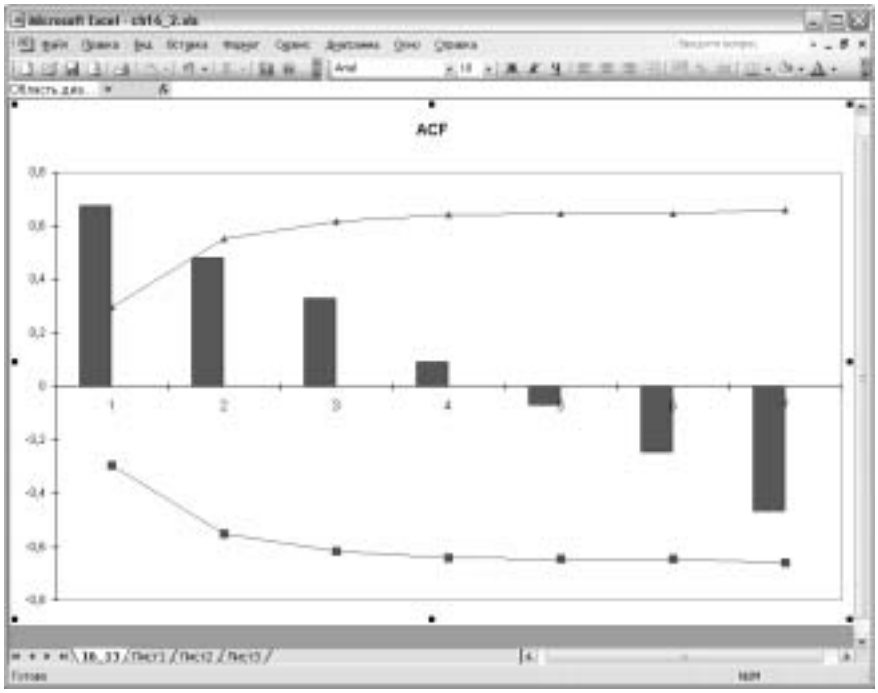


Рис. 16.13. Высота столбца соответствует уровню автокорреляции

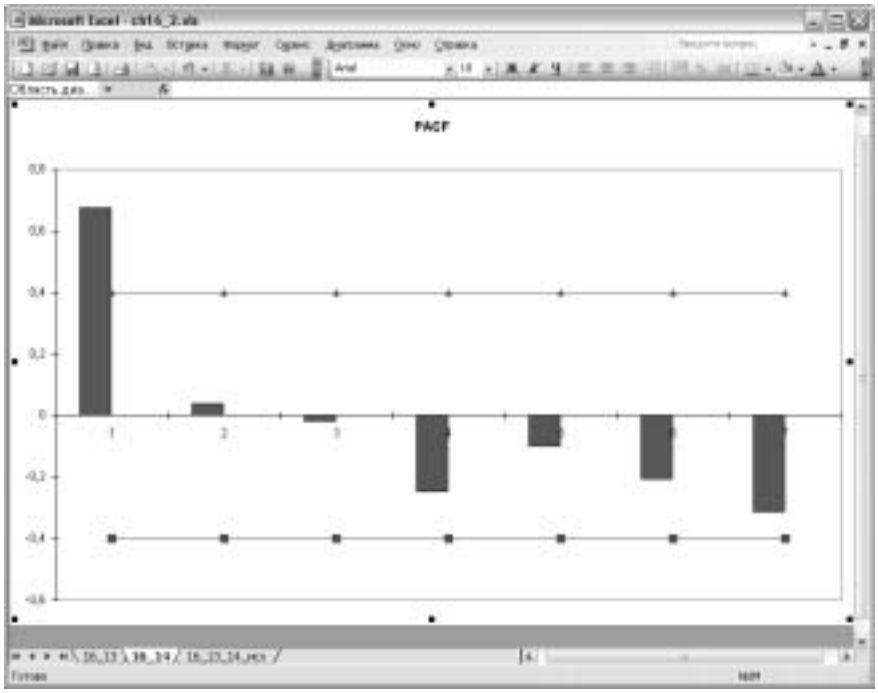


Рис. 16.14. Стационарный набор данных на диаграмме PACF обычно имеет один пик

Если бы пик на графике PACF достигался во втором столбике, при создании прогноза нужно было бы смещаться на два периода назад, т.е. для прогноза в диапазоне A4:A26 использовать значения диапазона A2:A24.

Регрессия одного тренда на другой

Если термины *причинность* и *корреляция* вам знакомы, в один прекрасный момент вы могли читать о различиях между этими понятиями. Давайте предположим, что некоторый ученый решил исследовать зависимость в 100 разных сообществах между количеством книг в публичных библиотеках и числом жителей, которые сделали карьеру в области космических исследований.

Предположим, что этот социолог нашел сильную зависимость между данными двумя переменными. Значит ли это, что если увеличить количество книг в публичных библиотеках, то количество выпускников школ, успешно поступивших в вузы, также возрастет?

Естественно, нет. Здесь нет причинности — только корреляция. Возможно, в этом вопросе причинностью обладает некоторая другая переменная (или несколько переменных), которую социолог упустил из виду. Сообщества с более высоким уровнем доходов имеют возможность содержать более богатые публичные библиотеки, а семьи, живущие в таких сообществах, имеют больше средств для поддержки высшего образования своих детей.

Если найти какой-либо способ повысить уровень доходов в каком-либо из сообществ, можно будет наблюдать как рост числа книг в публичных библиотеках, так и более высокий уровень образования. В этом и состоит суть метода поиска причинности: выбрать случайным образом некоторую группу людей, отделить ее от остальных, подвергнуть особой заботе и сравнить ее результаты с остальными.

Тот же эффект можно наблюдать при использовании регрессии в исследовании взаимосвязи между двумя переменными: прогнозируемой и той, которая каким-либо образом может быть связана с данной. Предположим, что мы располагаем набором данных о продажах, в которых в течение последних десяти лет наблюдался сильный повышательный тренд. В этом случае наверняка окажется, что в течение этого периода времени число товарных линий, предлагаемых компанией, также постоянно росло.

Из этого можно сделать следующий вывод: повысить объем продаж можно за счет расширения товарных линий. Однако число товарных линий и объем продаж могут одновременно расти и не будучи непосредственно связанными друг с другом. Товарные линии могут расширяться за счет влияния технологического прогресса, а рост объема продаж может быть связан с благоприятной средой бизнеса — тем обстоятельством, которое вызывает рост одних компаний и крах других.

В данном примере две переменные росли независимо одна от другой — они изменялись во времени. В процесс вмешалась третья переменная, от которой зависели две исследуемые. Также и в рассмотренном ранее примере количество книг в библиотеке и уровень образования зависели от уровня доходов населения, а не друг от друга.

Выход в данной ситуации состоит в следующем. Нужно нивелировать тренды в обеих переменных и посмотреть, удастся ли получить достоверное уравнение регрессии. Существует несколько методов нивелирования трендов, однако наиболее удобным можно назвать метод вычисления первых конечных разностей (см. главу 17).