**Построение модели множественной регрессии**

Исследуя модели [простой](http://baguzin.ru/wp/?p=6078) и [множественной](http://baguzin.ru/wp/?p=6136) регрессии, предполагалось, что зависимость между откликом *Y* и каждой из объясняющих переменных является линейной. Однако существуют и другие виды взаимосвязи. Одной из наиболее распространенных нелинейных взаимосвязей между двумя переменными является квадратичная зависимость. Для ее анализа предназначена модель квадратичной регрессии.[[1]](#footnote-1)

Материал будет проиллюстрирован сквозным примером: **прогнозирование продолжительности простоя художников, входящих в профсоюз.** Представьте себе, что вы — директор телевизионной станции и стремитесь сократить производственные расходы. В частности, художники, входящие в профсоюз, получают почасовую оплату, даже когда они ничего не делают. Эти часы называют часами простоя. Считается, что общее количество часов простоя за неделю зависит от общего количества времени, проведенного в офисе, общего количества часов, проведенных на выезде, времени, затраченного на озвучивание, и общей продолжительности работы. Постройте модель множественной регрессии, позволяющую наиболее точно предсказать количество часов простоя. Она позволит выявить причины возникающих простоев и уменьшить их количество в будущем. Как построить наиболее подходящую модель? С чего начать?

Модель квадратичной регрессии:

*(1) Yi = β0 + β1 + β2 + εi*

где β0 — сдвиг, β1 — коэффициент линейного эффекта, β2 — коэффициент квадратичного эффекта, εi – случайная ошибка переменной *Y* в *i*-ом наблюдении.

Модель квадратичной регрессии похожа на модель множественной регрессии с двумя переменными, за исключением того, что вторая объясняющая переменная является квадратом первой. Как и в модели множественной регрессии, выборочные коэффициенты регрессии *b0*, *b1* и *b2* представляют собой оценки параметров генеральной совокупности β0, β1 и β2. Таким образом, можно сформулировать следующую квадратичную модель с одной объясняющей переменной *Х1* и зависимой переменной *Y* (уравнение квадратичной регрессии):

*(2) Ŷi = b0 + b1 + b2*

где коэффициент *b0* является сдвигом, коэффициент *b1* оценивает линейный эффект, а коэффициент *b2* — квадратичный эффект.

**Вычисление коэффициентов регрессии и предсказание отклика.** Проиллюстрируем применение квадратичной модели на примере эксперимента, в котором изучается влияние зольной пыли на прочность бетона. Для этого была создана выборка, состоящая из 18 образцов 28-дневного бетона, прочность которого равна 4000 фунтов на дюйм. Объем зольной пыли колебался от 0 до 60%. Уровень значимости α = 0,05 (рис. 1).



Рис. 1. Прочность 28-дневного бетона и содержание зольной пыли в 18 образцах

Для того чтобы выбрать наиболее подходящую модель, описывающую зависимость прочности бетона от процента зольной пыли, построим диаграмму разброса (рис. 2). Как видим, при возрастании процента зольной пыли прочность бетона увеличивается, достигает максимума при содержании зольной пыли, равном 40%, а затем уменьшается. Итак, квадратичная модель точнее описывает исследуемую зависимость, чем линейная.

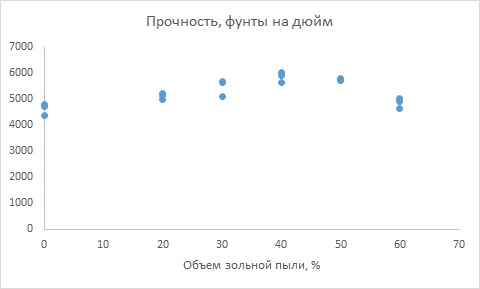


Рис. 2. Диаграмма разброса содержания зольной пыли (ось *X*) и прочности бетона (ось *Y*)

Значения трех коэффициентов регрессии (*b0*, *b1* и *b2*) можно вычислить с помощью *Пакета анализа* Excel. Предварительно нужно создать еще одну колонку со значениями *Х2* (рис. 3).

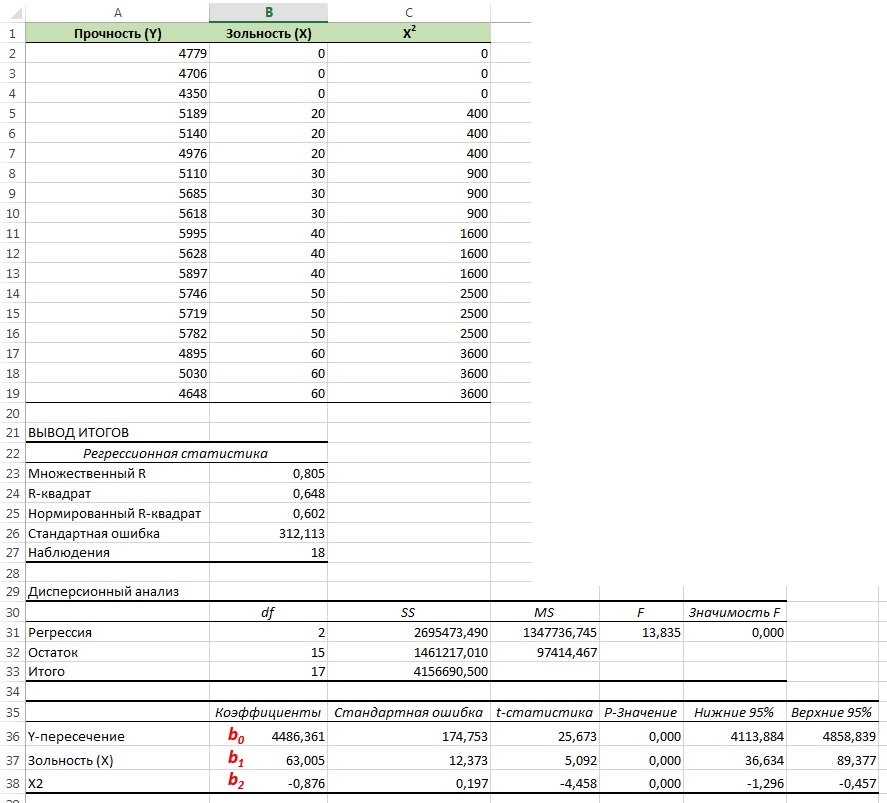


Рис. 3. Результаты регрессионного анализа, полученные с помощью *Пакета анализа* Excel при решении задачи о прочности бетона

Уравнение квадратичной регрессии имеет следующий вид:

*Ŷi = b0 + b1 + b2 = 4486,361 + 63,005 – 0,876*

где *Ŷi* — предсказанная прочность *i*-го образца, *Х1i* — содержание зольной пыли в *i*-ом образце.

Для того чтобы продемонстрировать соответствие построенной модели исходным данным, на рис. 4 приведен график квадратичной зависимости прочности бетона от содержания зольной пыли. Для построения графика нужно вернуться к рис. 2, кликнуть правой кнопкой мыши на точках диаграммы, и выбрать *Добавить линию тренда*. В открывшемся окне выбрать параметр линии тренда *Полиномиальная*, степень 2, а также кликнуть *Показывать уравнение на диаграмме*.

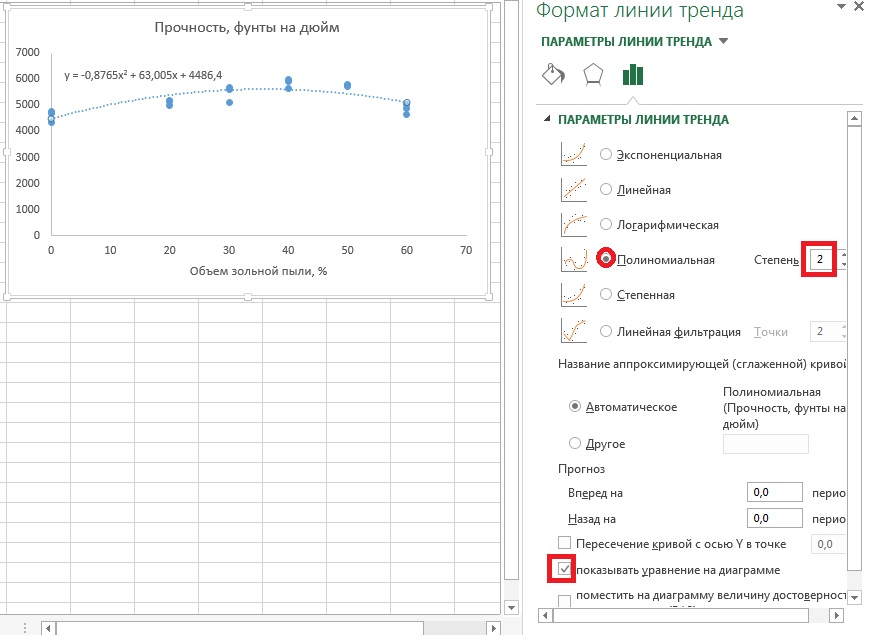


Рис. 4. График квадратичной зависимости на диаграмме разброса содержания зольной пыли (ось *X*) и прочности бетона (ось *Y*)

Коэффициент *b0*, представляющий собой предсказанную среднюю прочность бетона при нулевом содержании зольной пыли, представляет собой сдвиг отклика и равен 4486,361. Чтобы объяснить смысл коэффициентов *b1* и *b2*, следует обратить внимание на рис. 4. Как видим, при увеличении содержания зольной пыли прочность бетона сначала увеличивается, а затем уменьшается. Этот эффект можно продемонстрировать, предсказав среднюю прочность бетона при содержании зольной пыли, равном 20, 40 и 60%. Используя квадратичную модель:

*Ŷi = 4486,361 + 63,005 – 0,876*

получаем следующие результаты (рис. 5):

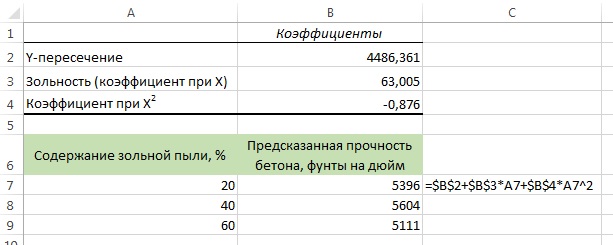


Рис. 5. Предсказанная прочность бетона на основе квадратичной модели

**Проверка значимости квадратичной модели.** Убедившись, что квадратичная модель адекватна исходным данным, можно проверить, существует ли статистически значимая зависимость между прочностью бетона *Y* и содержанием зольной пыли *X*. Нулевая и альтернативная гипотезы формулируются следующим образом: Н0: β1 = β2 = 0 (между откликом *Y* и объясняющей переменной *Х1* нет зависимости); Н1 : β1 ≠ 0 и/или β2 ≠ 0 (между откликом *Y* и объясняющей переменной *Х1* есть зависимость). Нулевую гипотезу можно проверить с помощью *F*-критерия:

*F = 13,835*

(см. рис. 3, ячейки D31, D32, Е31)

Если уровень значимости α = 0,05, критическое значение *F*-распределения, имеющего две и 15 степеней свободы, =F.ОБР(0,95;2;15) = 3,682 (рис. 6). Поскольку *F* = 13,84 > *FU* = 3,68 и *р* =1-F.РАСП(E31;2;15;ИСТИНА) = 0,00039 < 0,05, нулевая гипотеза *Н0* отклоняется. Таким образом, между прочностью бетона и содержанием зольной пыли существует статистически значимая зависимость.

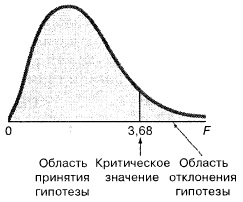


Рис. 6. Проверка гипотезы о существовании зависимости между откликом и объясняющей переменной, если уровень значимости равен 0,05, а *F*-распределение имеет две степени свободы в числителе и 15 – в знаменателе

**Оценка квадратичного эффекта.** Регрессионная модель, описывающая зависимость между двумя переменными, должна быть не только как можно более точной, но и максимально простой. Следовательно, необходимо проверить, существуют ли статистически значимые различия между квадратичной *Yi = β0 + β1 + β 2*и линейной *Yi = β0 + β1 + ε* моделями. Напомним, что для оценки вклада каждой поясняющей переменной используется *t*-критерий. Среднеквадратичная ошибка каждого коэффициента регрессии и соответствующие значения *t*-статистики приведены на рис. 3. Чтобы проверить значимость квадратичного эффекта, сформулируем следующую нулевую и альтернативную гипотезы: *Н0* — включение квадратичного эффекта не приводит к значительному увеличению точности модели (β2 = 0), *Н1* — включение квадратичного эффекта значительно повышает точность модели (β2 ≠ 0). *t*-статистика квадратичного эффекта (β2) = –4,458 (см. рис. 3, ячейка D38). Критические значения *t*-статистики, имеющего 15 степеней свободы при уровне значимости α = 0,05: *tL* =СТЬЮДЕНТ.ОБР(0,025;15) = –2,1315; *tU* =СТЬЮДЕНТ.ОБР(0,975;15) = +2,1315 (рис. 7).

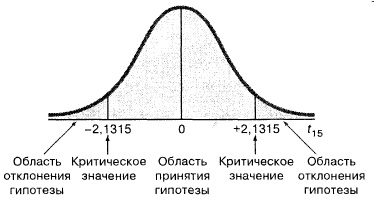


Рис. 7. Проверка гипотезы о вкладе квадратичного эффекта, если уровень значимости α = 0,05, a *t*-распределение имеет 15 степеней свободы

Поскольку *t* = –4,458 < *tL* = –2,1315, и, кроме того, *р* = 0,00046 < 0,05, нулевая гипотеза *Н0* отклоняется. Следовательно, квадратичный эффект значительно повышает точность предсказания по сравнению с линейной моделью, описывающей зависимость между прочностью бетона и содержанием зольной пыли.

**Коэффициент множественной смешанной корреляции** в модели множественной регрессии позволяет оценить долю вариации переменной *Y*, объясняемой изменениями двух объясняющих переменных. В квадратичном регрессионном анализе влияния содержания золы на прочность бетона этот коэффициент задается формулой:



В нашем примере SSR = 2 695 473 (рис. 3, ячейка С31), SST = 4 156 690 (ячейка С33). Таким образом, *rY.122* = 0,6485. Эта величина означает, что 64,85% вариации прочности бетона можно объяснить квадратичной зависимостью между прочностью бетона и содержанием зольной пыли.

**Преобразование данных в регрессионных моделях**

Перейдем к изучению регрессионных моделей, в которых независимая переменная *X*, зависимая переменная *Y* или обе переменные подвергаются преобразованиям, чтобы преодолеть ограничения, наложенные на модель, либо для ее линеаризации. К наиболее распространенным преобразованиям относятся извлечение квадратного корня или логарифмирование.

**Извлечение квадратного корня.** Для преодоления ограничений, связанных со свойством гомоскедастичности,[[2]](#footnote-2) а также для превращения нелинейной модели в линейную часто применяется извлечение квадратного корня. Если из объясняющей переменной извлекается квадратный корень, регрессионная модель принимает следующий вид:

*(3) Yi = β0 + β1 + εi*

*Пример 1.* Извлечение квадратного корня из переменной X (рис. 8а) превращает нелинейную зависимость (рис. 8б) в линейную (рис. 8в).

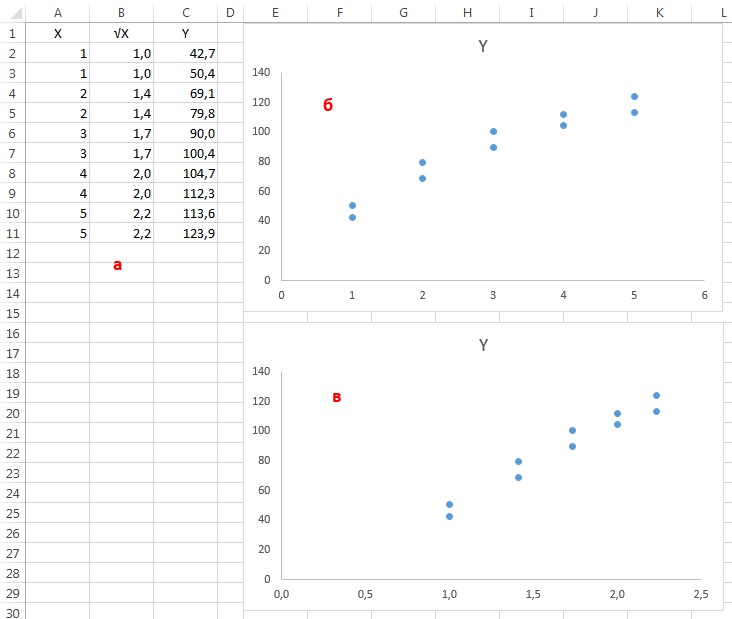


Рис. 8. Диаграммы разброса: (б) для исходных данных; (в) для квадратного корня из переменной X

**Логарифмическое преобразование.** Когда нарушается условие гомоскедастичности, кроме извлечения квадратного корня, часто применяется логарифмическое преобразование. Оно также позволяет превратить нелинейную модель в линейную. Чтобы не углубляться в сложные формулы, проиллюстрируем применение логарифмического преобразования на примере.

Пример 2. Диаграмма разброса (рис. 9а), демонстрирующая экспоненциальный рост исходных данных, может принять вид линейной путем преобразования зависимой и объясняющей переменных (рис. 9б). Удобнее всего это сделать простым выбором *Логарифмической шкалы* по обеим осям (рис. 9в). Иногда достаточно изменить только одну ось.

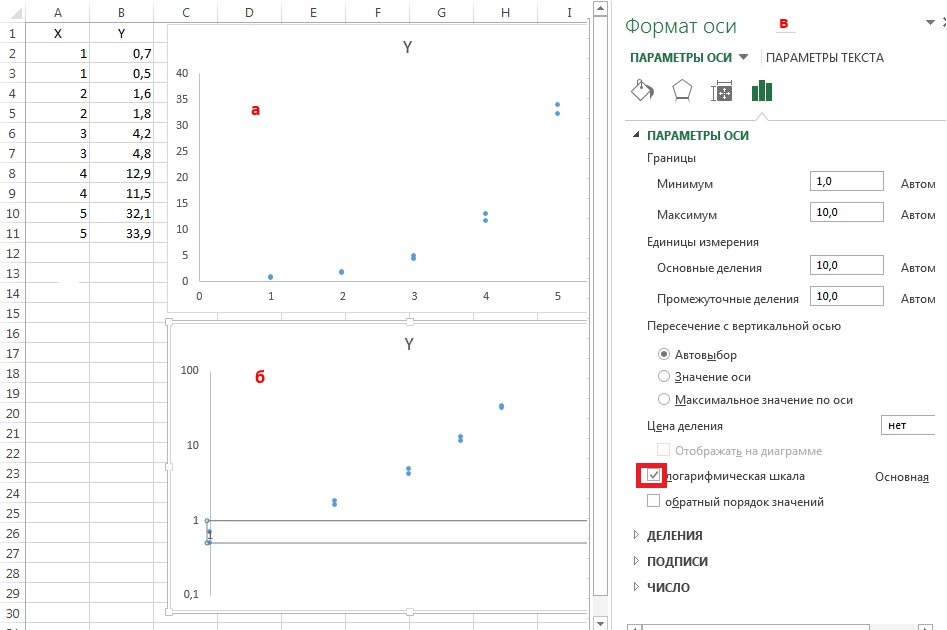


Рис. 9. Диаграммы разброса: (а) для исходных данных; (б) после логарифмического преобразования переменных *X* и *Y*; (в) показано, что преобразованы не исходные данные, а вид шкал на диаграмме

**Коллинеарность**

Применение модели множественной регрессии сопряжено с весьма важной проблемой — возможной коллинеарностью объясняющих переменных. Коллинеарными называют объясняющие переменные, значительно коррелирующие друг с другом. В этих ситуациях переменные не добавляют новой информации, поэтому их влияние на отклик трудно оценить. Это может привести к явной неустойчивости регрессионных коэффициентов, соответствующих коллинеарным переменным. Оценить коллинеарность можно, вычислив коэффициент инфляции (variance inflationary factor – VIF) для каждой объясняющей переменной. Коэффициент инфляции:

*(4) VIF =*

где *Rj2* — коэффициент множественной смешанной корреляции объясняющей переменной *Xj* со всеми другими объясняющими переменными.

Если модель содержит только две объясняющие переменные, величина *R12* представляет собой коэффициент смешанной корреляции между переменными *X1* и *Х2*. Он может совпадать с величиной *R22* — коэффициентом смешанной корреляции между переменными *Х2* и *Х1*. Если в модели содержатся три объясняющие переменные, то величина *Rj2*, где *j* = 1, 2, 3, представляет собой коэффициент множественной смешанной корреляции между переменной *Xj* и двумя другими объясняющими переменными.

Если объясняющие переменные не коррелируют друг с другом, коэффициент *VIFj* равен 1. Если объясняющие переменные сильно коррелируют друг с другом, *VIFj* может быть больше 10.

Модель множественной регрессии, в которой существуют большие коэффициенты инфляции, следует применять с крайней осторожностью. Эти модели позволяют предсказывать значения зависимой переменной только в том случае, если значения независимых переменных, подставляемые в модель, хорошо согласуются с данными, содержащимися в исходном наборе данных. Эти модели нельзя применять для экстраполяции отклика на значения независимых переменных, не содержащихся в исходной выборке. Кроме того, коэффициенты таких моделей не поддаются интерпретации, поскольку независимые переменные содержат перекрывающуюся информацию, а их индивидуальный вклад невозможно вычислить точно. Для решения этой проблемы следует исключить из регрессионной модели переменную, имеющую наибольший коэффициент инфляции. Довольно часто после этой операции сокращенная модель уже не содержит коллинеарных переменных.

Если вернуться к задаче о продажах батончиков OmniPower, рассмотренной [ранее](http://baguzin.ru/wp/?p=6136), окажется, что коэффициент корреляции между двумя объясняющими переменными (ценой и затратами на рекламу) равен –0,0968. Коэффициент инфляции этих переменных:

*VIF1 = VIF2 = = 1,009*

Таким образом, объясняющие переменные в задаче о продажах батончиков OmniPower не коллинеарны.

**Построение модели**

Остановимся подробнее на процессе построения модели, содержащей несколько объясняющих переменных. Для начала вспомним о задаче, в которой для предсказания объема простоя на телевизионной станции были учтены четыре объясняющие переменные (продолжительность работы в офисе, количество часов, проведенных на выезде, время, затраченное на озвучивание, и общее количество рабочих часов в неделе). Попробуем предсказать количество часов простоя, используя данные, приведенные на рис. 10.

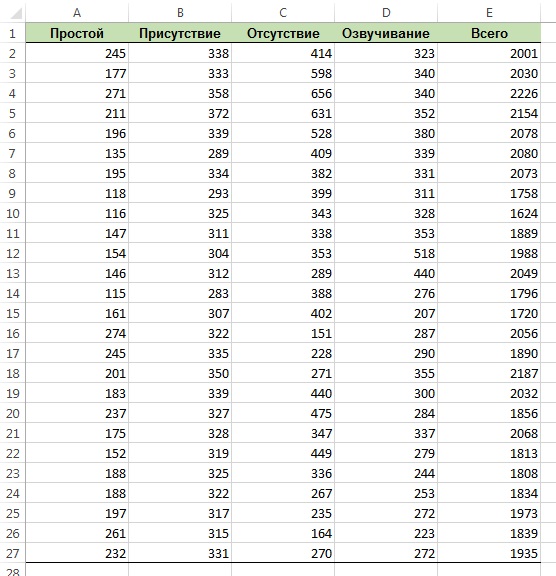


Рис. 10. Предсказание продолжительности простоя по количеству часов, проведенных в офисе, количеству часов, проведенных на выезде, количеству часов, затраченных на озвучивание, и общему количеству рабочих часов в неделе.

Прежде чем приступать к прогнозированию, необходимо учесть, что модель должна быть экономной. Это значит, что наша цель — разработать регрессионную модель, включающую в себя как можно меньше объясняющих переменных, позволяющих адекватно интерпретировать интересующий нас отклик. Регрессионная модель с минимальным количеством переменных намного проще других и меньше страдает от коллинеарности переменных. Кроме того, необходимо понимать, что модель с б*о*льшим количеством объясняющих переменных порождает б*о*льшие сложности при регрессионном анализе. Во-первых, оценка всех возможных регрессионных моделей становится крайне сложной вычислительной задачей. Во-вторых, даже если конкурентные модели удалось оценить, может оказаться, что единственной оптимальной модели не существует, а есть несколько одинаково хороших.

Начнем анализ простоев на телевизионной станции с оценки коллинеарности других объясняющих переменных, вычислив коэффициент инфляции (4) для каждой из них (рис. 11). Для этого необходимо исключить колонку *Простой*, а затем провести регрессионный анализ последовательно назначая в качестве зависимой переменной *Присутствие*, *Отсутствие*, *Озвучивание* и *Всего*, а в качестве объясняющих – три оставшиеся (подробнее см. Excel-файл).

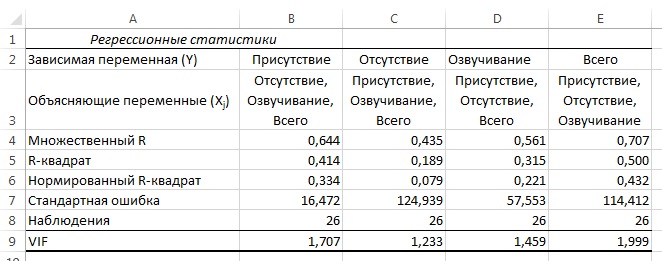


Рис. 11. Анализ коллинеарности объясняющих переменных

Обратите внимание на то, что коэффициенты VIF относительно малы и колеблются от 1,23 для часов, проведенных на выезде, до 2,0 для общего количества рабочих часов. Таким образом, поскольку коэффициенты VIF не больше пяти, мы можем утверждать, что объясняющие переменные не коллинеарны.

**Пошаговый подход к построению регрессионной модели.** Продолжим анализ задачи о простоях и попробуем определить такой набор объясняющих переменных, который позволил бы построить адекватную и точную модель без необходимости учитывать все переменные. Одним из основных способов построения таких моделей является пошаговая регрессия, с помощью которой можно определить наилучшую регрессионную модель без перебора всех регрессионных моделей. После определения наилучшей модели для проверки проводится анализ остатков.

Напомним, что для оценки вклада переменных в модель множественной регрессии применяется *F*-критерий. В процессе шаговой регрессии *F*-критерий применяется к модели с любым количеством переменных. Важным свойством пошаговой процедуры является то, что объясняющие переменные, включенные в модель на предыдущих этапах, могут впоследствии исключаться из рассмотрения. Это значит, что на каждом этапе объясняющие переменные как включаются, так и исключаются из модели. Пошаговая регрессия останавливается, когда ни добавление, ни удаление объясняющих переменных не повышают точность модели.

При включении объясняющих переменных в модель и удалении их из нее уровень значимости α принимается равным 0,05. Начнем с попарного анализа, в котором зависимой переменной является *Простой*, а объясняющей переменной (единственной) последовательно: *Присутствие*, *Отсутствие*, *Озвучивание* и *Всего* (рис. 12). Видно, что наиболее сильно коррелирует с откликом *Присутствие*. Поскольку *р*-значение равно 0,001 и меньше 0,05, эта переменная включается в регрессионную модель.

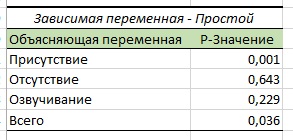


Рис. 12. Анализ влияния первой объясняющей переменной на отклик

На следующем этапе в модель включается вторая объясняющая переменная. Она должна иметь наибольшее влияние на точность модели при условии, что первая объясняющая переменная (продолжительность работы в офисе) уже учтена. В данной задаче такой переменной оказалось количество часов, проведенных на выезде (рис. 13). Поскольку *р*-значение, соответствующее этой переменной, равно 0,027 и не больше 0,05, количество часов, проведенных на выезде (отсутствие), включается в модель.

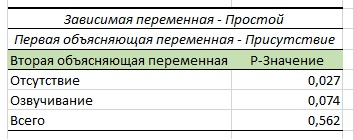


Рис. 13. Анализ влияния второй объясняющей переменной при условии, что первая объясняющая переменная (*Присутствие*) уже учтена

Теперь необходимо определить, насколько велик вклад продолжительности работы в офисе и не следует ли исключить его из модели. Поскольку *р*-значение для этой переменной равно 0,0001, ее следует оставить в модели (см. Excel-файл).

На следующем этапе необходимо решить, стоит ли включать в модель третью переменную (рис. 14). Поскольку ни одна из оставшихся переменных не удовлетворяет *F*-критерию с 5%-ным уровнем значимости, в результате получаем регрессионную модель с двумя объясняющими переменными: продолжительностью работы в офисе (присутствие) и количеством часов, проведенных на выезде (отсутствие).

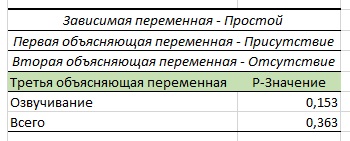


Рис. 14. Анализ влияния третьей объясняющей переменной при условии, что две объясняющие переменные (*Присутствие* и *Отсутствие*) уже учтены

Процедура пошаговой регрессии была предложена около тридцати лет назад, когда стоимость компьютерного времени была очень высока. В этих условиях она позволяла сократить объем перебора объясняющих переменных и широко использовалась. В настоящее время появились новые очень эффективные регрессионные модели. Так был разработан более общий подход к построению альтернативных регрессионных моделей, получивший название метода выбора наилучшего подмножества. В последнее время появилась новая методика исследования — интеллектуальный анализ данных — способ анализа информации в огромных базах данных для поиска статистически значимых зависимостей среди огромного количества объясняющих переменных. В этих условиях метод выбора наилучшего подмножества становится непрактичным.

С помощью **метода выбора наилучшего подмножества** либо оценивают всевозможные регрессионные модели для заданного набора данных, либо определяют наилучшие подмножества моделей для заданного количества независимых переменных. На рис. 15 показаны результаты применения метода выбора наилучшего подмножества для решения задачи о простоях на телевизионной станции. Обратите внимание на то, что максимальным значением скорректированного коэффициента *r2* является число 0,551. Оно достигается для модели, в которой учитываются четыре объясняющие переменные и эффект взаимодействия всех пяти оцениваемых параметров.

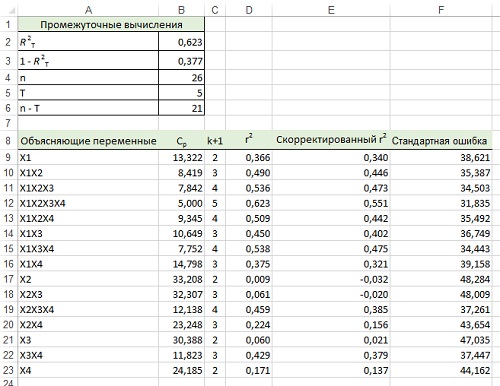


Рис. 15. Результаты применения метода выбора наилучшего подмножества для решения задачи о простоях на телевизионной станции; чтобы создать эту таблицу нужно последовательно провести регрессионный анализ для каждого набора объясняющих переменных (всего 15 раз, подробнее см. файл *Данные для построения рисунка 15.xlsx*); обратите внимание на чрезвычайно маленькое значение коэффициента *r2* и учтите, что скорректированный коэффициент *r2* может быть отрицательным.

В качестве второго критерия часто используется статистика, предложенная Мэллоусом. Статистика *Ср* оценивает разность между эмпирической и *истинной* регрессионной моделями:

где *n* – количество наблюдений (в нашем случае 26, см. рис. 10), *k* — количество независимых переменных, включенных в регрессионную модель, *Т* — общее количество параметров (включая эффекты взаимодействия), включенных в полную модель регрессии (*T = kmax + 1*),  — коэффициент множественной смешанной корреляции в регрессионной модели, содержащей *k* независимых переменных,  — коэффициент множественной смешанной корреляции в полной регрессионной модели, содержащей все *Т* оцениваемых параметра.

Вычислим статистику *Ср* для модели, содержащей продолжительность работы в офисе и количество часов, проведенных на выезде, используя вышеприведенную формулу:

*n* = 26, *k* = 2, *T* = 4 + 1 = 5, = 0,490,  = 0,623.

Таким образом,

Если отклонения регрессионной модели, содержащей *k* независимых переменных, от истинной модели являются случайными, среднее значение статистики *Ср* равно *k + 1*, т.е. количеству параметров. Таким образом, при оценке многих альтернативных регрессионных моделей основная цель — найти модели, для которых величина *Ср* близка *k + 1* или меньше этого числа. Как показано на рис. 15, этому критерию соответствует лишь одна модель, содержащая все четыре независимые переменные. Следовательно, необходимо выбрать именно эту модель. Довольно часто статистика *Ср* выделяет не одну, как в данном случае, а несколько моделей, которые подлежат более глубокому анализу на основе критериев экономии, простоты и соответствия исходным предположениям (по результатам анализа остатков). Обратите также внимание на то, что значение статистики Ср для модели, выбранной по результатам пошагового анализа, равно 8,4. Эта величина намного превышает предполагаемый уровень *k* + 1 =3.

Определив объясняющие переменные, которые следует включить в модель, необходимо проверить ее точность с помощью анализа остатков (рис. 16). Обратите внимание на то, что все графики не демонстрируют никаких явных зависимостей.

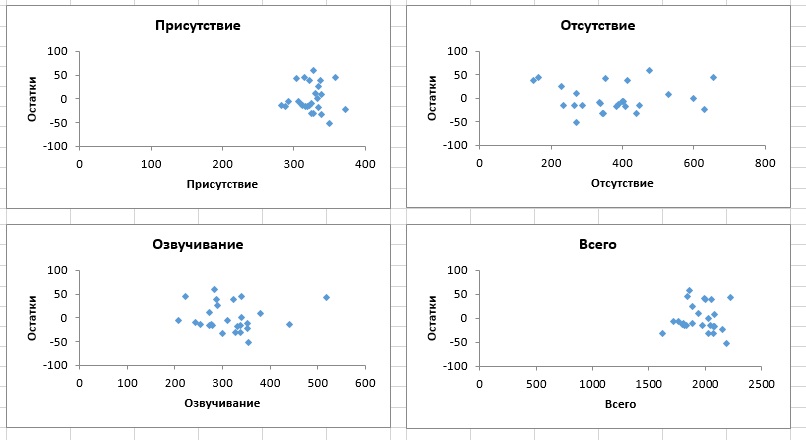


Рис. 16. Графики остатков, построенные с помощью Пакета анализа Excel при решении задачи о простоях

Этапы построения регрессионной модели (рис. 17):

1. Определить набор независимых переменных для включения в регрессионную модель.
2. Построить полную регрессионную модель, учитывающую все независимые переменные, и вычислить коэффициент VIF для каждой из них.
3. Определить, все ли независимые переменные имеют коэффициент VIF больше пяти.
4. Возможны три варианта:
   * Для всех независимых переменных коэффициент VIF больше пяти. Перейти к п. 5.
   * Для одной независимой переменной коэффициент VIF больше пяти. Исключить ее из модели и, перейти к п. 5.
   * Для нескольких независимых переменных коэффициент VIF больше пяти. Исключить из модели независимую переменную, имеющую наибольший коэффициент VIF, и перейти к п. 2.
5. Применить метод выбора наилучшего подмножества к оставшимся переменным и определить наилучшую модель (по величине *Ср*).
6. Перечислить все модели, у которых *Ср* ≤ k + 1.
7. Выбрать среди моделей, обнаруженных в п. 6, наилучшую.
8. Выполнить полный анализ выбранной модели, включая анализ остатков.
9. В зависимости от результатов анализа остатков добавить квадратичные члены, преобразовать данные и выполнить повторный анализ.
10. Применить полученную модель, чтобы предсказать значения зависимой переменной.

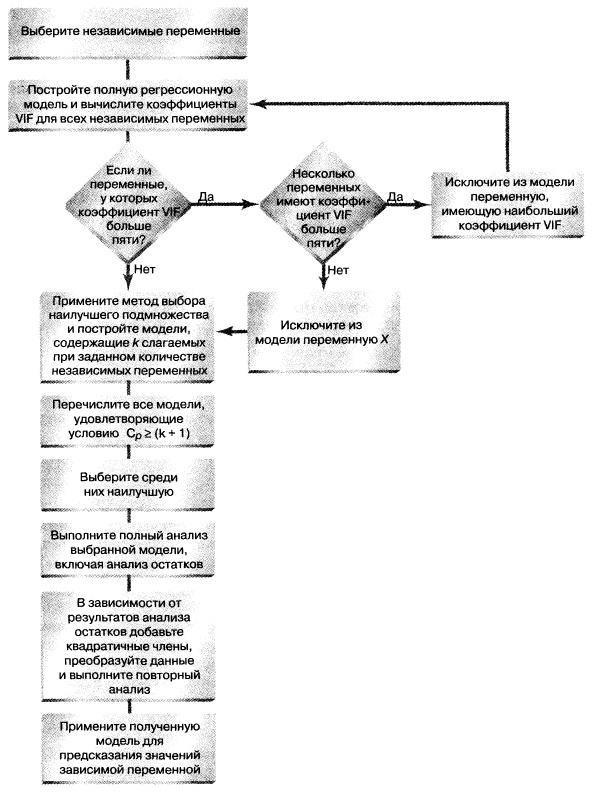


Рис. 17. Схема построения модели

**Ловушки и этические проблемы, связанные со множественной регрессией**

Построение моделей является синтезом искусства и науки. Разные люди придерживаются разных точек зрения на оптимальность регрессионных моделей. В любом случае рекомендуем придерживаться схемы на рис. 17. Однако применение этой схемы сопряжено с некоторыми ловушками:

* Необходимо понимать, что при интерпретации коэффициента регрессии, соответствующего конкретной независимой переменной, остальные переменные считаются константами.
* Следует проводить анализ остатков для каждой независимой переменной.
* Нужно оценивать эффект взаимодействия и проверять, чтобы наклоны отклика по каждой из объясняющей переменной были одинаковыми.
* Необходимо вычислять коэффициенты VIF для каждой независимой переменной, включаемой в модель.
* Следует проверять несколько альтернативных моделей, используя метод выбора наилучшего подмножества.

**Этические вопросы** возникают, когда модель множественной регрессии используется для предсказания величин, находящихся под управлением пользователя. Ключевым моментом в этом случае являются намерения исследователя. Возможны варианты, когда статистик преднамеренно не исключает из модели множественной регрессии коллинеарные переменные и неправомерно применяет метод наименьших квадратов даже тогда, когда не выполняются необходимые условия.

**Резюме.** В заметке показано, как директор телевизионной станции может применять множественный линейный анализ для сокращения продолжительности простоев. Рассмотрены различные модели множественной регрессии, включая квадратичные модели, модели с фиктивными переменными, модели с эффектами взаимодействия. Изучены способы преобразования переменных, исследованы коллинеарные переменные и описан процесс построения регрессионной модели.

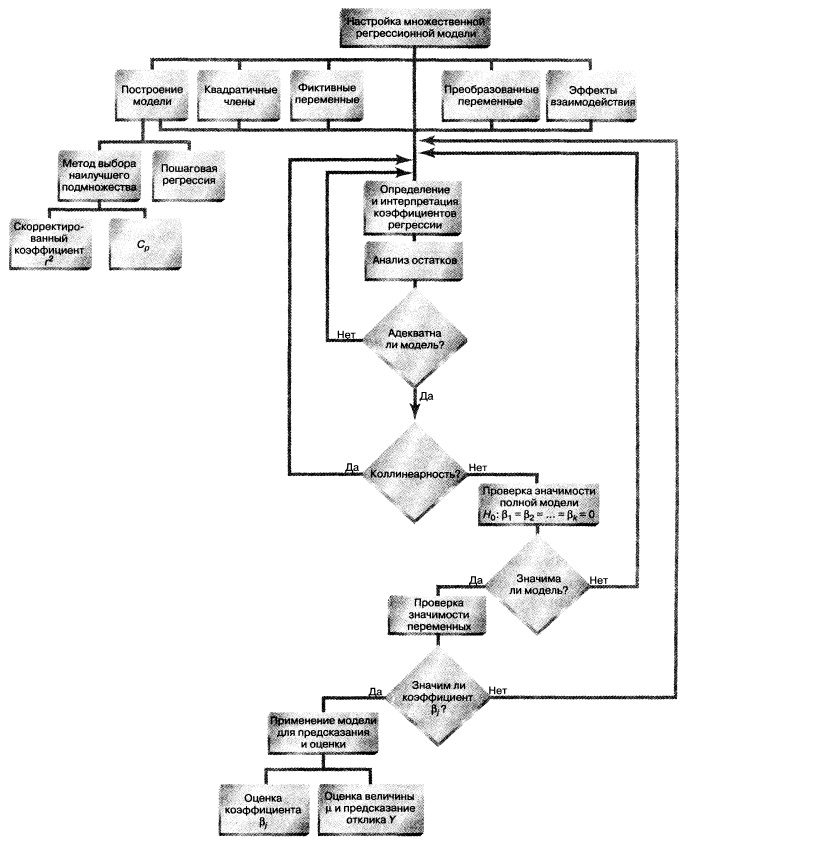


Рис. 18. Структурная схема заметки

Предыдущая заметка [Введение в множественную регрессию](http://baguzin.ru/wp/?p=6136)

Следующая заметка

К оглавлению [Статистика для менеджеров с использованием Microsoft Excel](http://baguzin.ru/wp/?p=5285)

1. Используются материалы книги Левин и др. Статистика для менеджеров. – М.: Вильямс, 2004. – с. 937–981 [↑](#footnote-ref-1)
2. [Гомоскедастичность](http://ru.wikipedia.org/wiki/%D0%A1%D0%BA%D0%B5%D0%B4%D0%B0%D1%81%D1%82%D0%B8%D1%87%D0%BD%D0%BE%D1%81%D1%82%D1%8C) – равенство дисперсий случайных отклонений для различных *Х*, то есть, распределение предсказанного отклика *Y* вокруг среднего значения *Y̅* одинаково для всех *Х*. [↑](#footnote-ref-2)