Специальный математический практикум

с.н.с. Уфимцев Михаил Валентинович
факультет ВМК МГУ

Множественный регрессионный анализ и метод наименьших квадратов

0. Обозначения и сокращения

Здесь приняты следующие условные обозначения и сокращения:

1. Векторы и матрицы набраны полужирным шрифтом; обычно матрицы обозначаются заглавными буквами, а векторы – прописными, причём над вектором стоит стрелка; например, $\boldsymbol{X}$ – матрица, а $\vec{y}$ - вектор.
2. Вектор параметров регрессионной модели обозначается как $\vec{\boldsymbol{\theta}}$.
3. Размерность векторов и матриц кратко обозначается как $\boldsymbol{A} \in R^{q\ x\ k}$ (действительная матрица $\boldsymbol{A}$ размеров $q\ x\ k$) и $\vec{c} \in R^{q}$ (вещественный вектор длины q).
4. Под вектором понимается вектор-<u>столбец</u>.
5. $\vec{\boldsymbol{0}}$ – вектор, состоящий из нулей; $\vec{\boldsymbol{1}}_{\boldsymbol{n}}$ – n-мерный вектор из единиц; $\boldsymbol{0}$ – матрица, состоящая из нулей; $\boldsymbol{I_n}$ – единичная матрица размера $n \times n$.
6. Запись типа $\vec{\boldsymbol{\varepsilon}} \sim N(\vec{\boldsymbol{0}}, \sigma^2 \boldsymbol{I_n})$ означает: "случайная величина $\vec{\boldsymbol{\varepsilon}}$ (в данном случае – вектор) имеет нормальное распределение с параметрами $\vec{\boldsymbol{0}}$ и $\sigma^2 \boldsymbol{I_n}$".
7. $E\{.\}$ – оператор взятия математического ожидания; $Var\{.\}$ – дисперсия; cov{. , .} – ковариационная матрица (или оператор ковариации); corr{. , .} – коэффициент корреляции.
8. Обычно для обозначения оценки используется символ " ^ " ("hat") над одноимённой буквой: например, если рассматривается вектор $\vec{\boldsymbol{\theta}}$, то его оценка – $\hat{\vec{\boldsymbol{\theta}}}$.
9. Используются следующие сокращения:
   а) МНК - метод наименьших квадратов;
   б) МП - максимального правдоподобия;
   в) м.о. - математическое ожидание;
   г) НК-оценка – оценка наименьших квадратов:
   д) с.в. – случайная величина;
   е) ф.п.в. – функция плотности вероятностей.

1. Постановка задачи

Рассматривается частный случай общей задачи описания связи между двумя переменными t и y: $y = \tilde{h}(t)$. Величины t и y предполагаются случайными. В регрессионном анализе эти переменные неравноправны: t – детерминированная величина (или случайная величина (с.в.), измеренная с гораздо более высокой точностью, чем y), а y – случайная величина, так что можно рассмотреть n измерений $t_1, \ldots, t_n$ (заметим, что <u>не</u> требуется, чтобы <u>все</u> точки измерений $t_1, \ldots, t_n$ были различны – см. далее).

Случайный характер y обусловлен ошибками измерений: в i-м наблюдении

$$y_i = \eta(t_i) + \varepsilon_i ,$$

где $\eta(t_i)$ – функция известного вида (например: 1) линейная; 2) полином (алгебраический); 3) тригонометрический полином; 4) сумма экспонент или функций Гаусса, и т.д.), определённая при каждом значении i аргумента t , а ε_i – случайная (т.е. $E\{\varepsilon_i\} = 0$) ошибка измерений.

Функция η известна не полностью: она содержит k неизвестных параметров $\vec{\theta} = (\theta_0, \ldots, \theta_{k-1})^T$, и цель регрессионного анализа – по наблюдениям $\vec{y} = (y_1, \ldots, y_n)^T$ оценить эти параметры, с указанием точности оценок. Форма зависимости η от $\theta_0, \ldots, \theta_{k-1}$ может быть нелинейной (нелинейный регрессионный анализ), а в частном случае – линейной (**множественный регрессионный анализ**). Требование аддитивности ошибок ε_i в y_i и случайности ε_i является главным как в линейном, так и в нелинейном регрессионном анализе.

В данной части практикума рассматриваются задачи множественного регрессионного анализа, т.е. зависимость y от параметров – линейная:

$$y_i = \theta_0 + \theta_1 g_1(t_i) + \ldots + \theta_{k-1} g_{k-1}(t_i) + \varepsilon_i, \qquad i = 1, 2, \ldots, n; \qquad (1)$$

первый параметр θ_0 соответствует "фиктивной" функции $g_0(t_i) \equiv 1$ и отражает постоянную часть в **функции регрессии** η. Вектор $\vec{y}$ наблюдений называется откликом, функции $g_0, g_1, \ldots, g_{k-1}$ *известны* (скажем, заданы аналитическими выражениями, определяются численно как решение системы дифференциальных уравнений, или заданы таблично <файлами>) и называются регрессорами (предикторными переменными).

Зарубежными исследователями подсчитано, что множественный регрессионный анализ – наиболее часто используемый статистический метод: на его долю приходится 70 - 80 % обращений к статистическим процедурам. Множественный регрессионный анализ хорошо изучен как теоретически, так и в аспекте эффективных и надёжных программных реализаций. Даже в Excel, Maple, Origin есть программы, выполняющие регрессионный анализ (в большей или меньшей степени). Для более полного ознакомления с ним можно рекомендовать имеющуюся на русском языке литературу [1–3]. Практическое овладение его методикой – цель данного практикума.

Из анализа вы знаете, что если $k > n$, то задача определения (оценивания) $\vec{\theta}$ по $\vec{y}$ неразрешима, т.е. нужно потребовать, чтобы выполнялось условие $k \le n$, причём в случае $k = n$ можно только оценить $\vec{\theta}$, но точность такой оценки будет неизвестна, если не считать заранее заданной стандартную ошибку в ε_i.

Модель (1) может быть записана более компактно в матричном виде:

$$\vec{y} = \boldsymbol{X}\,\vec{\theta} + \vec{\varepsilon}, \qquad (1')$$

где $\vec{y} = (y_1, \ldots, y_n)^T$, $\vec{\varepsilon} = (\varepsilon_1, \ldots, \varepsilon_n)^T$, $\vec{\theta} = (\theta_0, \ldots, \theta_{k-1})^T$, $\boldsymbol{X} = \| X_{ij} \| = \| g_j(t_i) \|$ $(i = 1, \ldots, n;\ j = 0, \ldots, k-1)$ – матрица данных размера $n \times k$ известна, отклик $\vec{y}$ известен, вектор ошибок $\vec{\varepsilon}$ ненаблюдаемый, вектор регрессионных параметров $\vec{\theta}$ неизвестен и подлежит оценке. Матрицу данных $\boldsymbol{X}$ можно записать как $\boldsymbol{X} = \| \boldsymbol{X}^{(1)} : \boldsymbol{X}^{(2)} : \ldots : \boldsymbol{X}^{(k)} \|$ – кокатенацию (сцепление) первого $\boldsymbol{X}^{(1)}$, второго $\boldsymbol{X}^{(2)}$ …, k-го $\boldsymbol{X}^{(k)}$ столбцов, где $\boldsymbol{X}^{(j)} \in R^n$ соответствует значениям j-го регрессора g_j в точках $t_1, t_2, \ldots, t_n$.

Для завершения постановки задачи в *классической модели* регрессионного анализа потребуем, чтобы

1) $E\{\ \vec{\varepsilon}\ \} = \vec{0}$ (случайность ошибок);
2) $\mathrm{cov}\{\ \vec{\varepsilon}\ ,\ \vec{\varepsilon}\ \} = E\{\vec{\varepsilon}\ \vec{\varepsilon}^{\,T}\} = \sigma^2 \boldsymbol{I_n}$ (некоррелированность ошибок и общая дисперсия их);
3) для $j = 0, \ldots, k-1$: $-\infty < \theta_j < \infty$ (рассматриваются значения параметров во всём k-мерном пространстве);
4) $rank(\boldsymbol{X}) = k$ (матрица данных имеет полный столбцовый ранг).

2. Основные результаты

По **методу наименьших квадратов (МНК)** (Гаусс, Лежандр) ищутся оценки параметров, минимизирующие сумму квадратов уклонений данных от значений, описываемых функцией регрессии (1), (1'). Иными словами, рассматривается задача минимизации квадратичной по θ_j целевой функции

$$Q(\vec{\theta}\ ) = (\vec{y}\ - \boldsymbol{X}\ \vec{\theta}\ )^T (\vec{y}\ - \boldsymbol{X}\ \vec{\theta}\ ) = \|\vec{y}\ - \boldsymbol{X}\ \vec{\theta}\ \|^2 = \sum_{i=1}^{n}\left(y_i - \sum_{j=0}^{k-1} X_{ij}\,\theta_j\right)^2; \qquad (2)$$

её решение всегда существует, единственно (в силу требования 4) и называется оценкой наименьших квадратов (НК-оценкой).

Минимизация (2) по $\vec{\theta}$ сводится (продифференцируйте $Q(\vec{\theta}\ )$ по $\vec{\theta}$ и приравняйте производную к $\vec{0}$) к решению системы k нормальных уравнений

$$\boldsymbol{X}^T\boldsymbol{X}\ \vec{\theta}\ = \boldsymbol{X}^T\vec{y}\ . \qquad (3)$$

Нетрудно показать [2, 3], что матрица $\boldsymbol{X}^T\boldsymbol{X}$ размера $k \times k$ – невырожденная, и поэтому НК-оценка есть

$$\hat{\vec{\theta}}\ = Arg\min_{\vec{\theta}} Q(\vec{\theta}) = (\boldsymbol{X}^T\boldsymbol{X})^{-1}\boldsymbol{X}^T\vec{y}\ . \qquad (4)$$

При выполнении условий 1 - 4 НК-оценка (а) является несмещённой, $E\{\ \hat{\vec{\theta}}\ \} = \vec{\theta}$; (б) её ковариационная матрица

$$\mathrm{cov}\{\hat{\vec{\theta}}\ ,\ \hat{\vec{\theta}}\ \} = E\{(\hat{\vec{\theta}} - \vec{\theta}\ )(\hat{\vec{\theta}} - \vec{\theta}\ )^T\} = \sigma^2 (\boldsymbol{X}^T\boldsymbol{X})^{-1}$$

всецело определяется матрицей данных $\boldsymbol{X}$ при любом законе распределения ошибок $\vec{\varepsilon}$; (в) она наилучшая в классе линейных несмещённых оценок, т.е. если $\tilde{\vec{\theta}}\ = \boldsymbol{A}\ \vec{y}$ – какая-то оценка с $E\{\ \tilde{\vec{\theta}}\ \} = \vec{\theta}$, то $\mathrm{cov}\{\tilde{\vec{\theta}}\ ,\ \tilde{\vec{\theta}}\ \} - \mathrm{cov}\{\hat{\vec{\theta}}\ ,\ \hat{\vec{\theta}}\ \} \geq \vec{0}$ и для $j = 0, \ldots, k-1$: $Var\{\ \hat{\theta}_j\ \} \leq Var\{\ \tilde{\theta}_j\ \}$.

Коль скоро $\hat{\vec{\theta}}$ вычислена по (4), можно подставить её в (1') и вычислить **прогноз (предсказание)**, даваемый рассматриваемой регрессионной моделью *не только при* рассматриваемом множестве значений аргумента $\{\ t_1, \ldots, t_n\ \}$, но и в произвольной точке t_* :

$$y_* = \hat{\theta}_0 + \hat{\theta}_1\ g_1(t_*) + \ldots + \hat{\theta}_{k-1}\ g_{k-1}(t_*)\ . \qquad (5)$$

В частности, если рассмотреть полное множество аргумента $\{\ t_1, \ldots, t_n\ \}$, то в матричном представлении получим вектор прогноза

$$\hat{\vec{y}}\ = \boldsymbol{X}\ \hat{\vec{\theta}}\ . \qquad (6)$$

Разность наблюдаемых и предсказываемых величин даёт **вектор регрессионных остатков**

$$\vec{e}\ = \vec{y}\ - \hat{\vec{y}}\ , \qquad (7)$$

а сумма квадратов e_i называется **остаточной суммой квадратов** <Residual Sum of Squares – ***RSS*** >,

$$RSS = ||\vec{e}\ ||^2 = \sum_{i=1}^{n} (y_i - \hat{y}_i)^2 \ , \qquad (8)$$

и она играет фундаментальную роль в регрессионном анализе. В частности, обычно в практических задачах общая дисперсия σ^2 неизвестна, и она оценивается по наблюдениям с помощью *RSS*:

$$\hat{\sigma}^2 = \frac{RSS}{n-k} = \frac{||\vec{y} - \hat{\vec{y}}||^2}{n-k} \ . \qquad (9)$$

В общем случае дисперсии ошибок ε_i различны, и ошибки могут быть коррелированы; при этом в условии 2 постановки задачи вместо $\sigma^2 \boldsymbol{I_n}$ будет фигурировать матрица $\boldsymbol{\Sigma}$ = cov{ $\vec{\varepsilon}$, $\vec{\varepsilon}$ }. Считая матрицу $\boldsymbol{\Sigma}$ известной и > $\boldsymbol{0}$, можно свести эту новую ситуацию к классической модели и показать, что *обобщённая НК-оценка* $\hat{\boldsymbol{\theta}}_g$ равна

$$\hat{\boldsymbol{\theta}}_g = (\boldsymbol{X}^T \boldsymbol{\Sigma}^{-1} \boldsymbol{X})^{-1} \boldsymbol{X}^T \boldsymbol{\Sigma}^{-1} \vec{\boldsymbol{y}} \ , \qquad (10)$$

несмещённая и имеет ковариационную матрицу

$$\mathrm{cov}\{ \hat{\boldsymbol{\theta}}_g , \hat{\boldsymbol{\theta}}_g \} = (\boldsymbol{X}^T \boldsymbol{\Sigma}^{-1} \boldsymbol{X})^{-1} . \qquad (11)$$

В частном случае диагональной $\boldsymbol{\Sigma}$ = diag{σ^2_1 , σ^2_2 , … , σ^2_n} <u>взвешенная</u> НК-оценка получается из минимизации целевой функции

$$Q_g(\vec{\boldsymbol{\theta}}\ ) = (\vec{\boldsymbol{y}} - \boldsymbol{X}\ \vec{\boldsymbol{\theta}}\ )^T \boldsymbol{\Sigma}^{-1} (\vec{\boldsymbol{y}} - \boldsymbol{X}\ \vec{\boldsymbol{\theta}}\ ) = \sum_{i=1}^{n} \frac{(y_i - \sum_j X_{ij}\, \theta_j)^2}{\sigma_i^2} \ , \qquad (3')$$

и легко сообразить, что можно привести задачу к целевой функции (3), если поделить на σ_i элементы X_{ij} i-ой строки матрицы данных, а также i-ый отклик y_i.

Вот каковы достижения метода наименьших квадратов в случае произвольного закона распределения ошибок. Свойство (в) оптимальности НК-оценок гарантирует их превосходство только в классе <u>линейных</u> оценок, а не всевозможных. Так, если оценивается единственный параметр θ_0 и если ошибки ε_i имеют распределение Лапласа, то [3] НК-оценка $\hat{\theta}_0 = \bar{y}$ (среднее значение) имеет только 50 % эффективности по отношению к наилучшей (максимального правдоподобия, МП-) оценке $\theta_{МП}$ = *выборочной медиане* { y_1 , …, y_n }.

Однако если ограничиться нормальным законом распределения ошибок (**нормальная регрессия**), как обычно предполагается в научно-технических измерениях, т.е. если

$$5)\ \vec{\varepsilon}\ \sim N(\vec{\boldsymbol{0}}\ , \sigma^2 \boldsymbol{I_n})$$

(или в обобщённой постановке с матрицей $\boldsymbol{\Sigma}$), то легко показать, что НК-оценка совпадает с МП-оценкой, т.е. является оптимальной в классе всевозможных оценок при $n \to \infty$.

В этом случае распределение $\hat{\boldsymbol{\theta}}$ нормальное:

$$\hat{\boldsymbol{\theta}}\ \sim N(\vec{\boldsymbol{\theta}}\ , \sigma^2 (\boldsymbol{X}^T \boldsymbol{X})^{-1}), \qquad (12)$$

распределение вектора остатков $\vec{e}$ – n-мерное вырожденное нормальное, а $\hat{\theta}$ и RSS статистически независимы.

В нормальной регрессии мы знаем не только распределение линейных функций $\vec{y}$ – векторов $\vec{e}$ и $\hat{\theta}$, но и некоторых квадратичных форм. Так, с.в. RSS/σ^2 имеет распределение χ^2_{n-k}:

$$\frac{RSS}{\sigma^2} \sim \chi^2_{n-k} , \qquad (13)$$

а

$$\frac{(\hat{\theta}-\vec{\theta})^T X^T X (\hat{\theta}-\vec{\theta})}{\sigma^2} \sim \chi^2_k . \qquad (14)$$

На утверждениях (12) - (14) основаны применения нормальной регрессии к построению доверительных областей для параметров и проверке общей линейной гипотезы H об ограничениях на $\vec{\theta}$ вида

$$A\,\vec{\theta} = \vec{c} , \qquad (15)$$

вектор $\vec{c} \in R^q$ $(q < k)$, и матрица $A \in R^{q\,x\,k}$ известны.

Так, если σ^2 неизвестна и оценена по (9), $\hat{\theta}$ – НК-оценка (4), а θ_j – истинное значение j-ой компоненты вектора $\vec{\theta}$, то с заданной вероятностью $\gamma = 1 - \alpha$ случайный интервал

$$\hat{\theta}_j - t_{n-k,\,1-\alpha} \sqrt{\hat{\sigma}^2 (X^T X)^{jj}} \le \theta_j \le \hat{\theta}_j + t_{n-k,\,1-\alpha} \sqrt{\hat{\sigma}^2 (X^T X)^{jj}} \qquad (16)$$

содержит θ_j. Здесь обозначено $(X^T X)^{jj} \equiv (X^T X)^{-1}{}_{jj}$, а $t_{n-k,\,1-\alpha}$ есть $(1-\alpha)$-квантиль распределения Стьюдента с $n-k$ степенями свободы (d.f.), т.е. если $f_{n-k}(z)$ – функция плотности вероятностей (ф.п.в.) с $n-k$ d.f., то $t_{n-k,\,1-\alpha}$ есть решение уравнения

$$1 - \alpha = \int_{-\infty}^{t_{n-k,\,1-\alpha}} f_{n-k}(z)\, dz . \qquad (17)$$

Можно найти доверительную область и для всего вектора $\vec{\theta}$ [1 - 3].

Задача проверки общей линейной гипотезы решается на основе *F-отношения*

$$F = \frac{(\vec{c} - A\hat{\theta})^T B^{-1} (\vec{c} - A\hat{\theta})}{q\,\hat{\sigma}^2} \qquad (18)$$

с матрицей $B = A\,(X^T X)^{-1} A^T$. Если H верна, то с.в. F имеет распределение Снедекора (F-распределение) с q, $n-k$ d.f. Критическими для проверки H являются <u>большие</u> значения F-отношения: если для выбранного уровня значимости α вычисленное $F > F_{q,\,n-k,\,1-\alpha}$, то гипотеза H отвергается (F-критерий). Критическое значение $F_{q,\,n-k,\,1-\alpha}$ имеет примерно тот же смысл, что и величина в (17), однако ф.п.в. распределения Снедекора не равна 0 только при положительных значениях z, так что нижний предел интегрирования равен 0.

В большинстве ситуаций рассматривается частный случай (15) с $\vec{c} = \vec{0}$, $A = [0 : I_q]$, матрица $0 \in R^{(k-q)\,x\,q}$ состоит из одних нулей. Иными словами, проверяется гипотеза H', что последние q компонент вектора $\vec{\theta}$ равны 0 (или некоторым фиксированным значениям c_l, $l = 1,\ldots, q$, но чаще всего предполагают $c_l = 0$). Это задача выбора между более простой (H' верна) и

более сложной вложенными моделями (базисные функции g_j для обеих моделей одинаковы).

Если же делают выбор из двух или нескольких не вложенных моделей, то обычно применяют неформальные графические методы – анализ регрессионных остатков [1, 2]. В принципе, они содержат всю статистическую информацию о качестве подгонки данных регрессионной моделью.

В заключение этого раздела – несколько слов о терминах "метод наименьших квадратов" и "регрессионный анализ". Первый касается преимущественно технической, вычислительной стороны вопроса: постановки задачи минимизации и её решения путём сведения к системе (3) нормальных уравнений. Это в основном – линейная алгебра, а статистики тут почти нет (для доказательства справедливости свойств (а) - (в) НК-оценок достаточно существования первых двух моментов вектора $\vec{\boldsymbol{\varepsilon}}$, а никаких дальнейших предположений о его распределении не делается). Термин "регрессионный анализ" фактически равнозначен "применению результатов нормальной регрессии" (построение доверительных интервалов и областей, проверка гипотез, анализ остатков). В нём можно получить больше, но и требуется больше.

3. Линейная одномерная регрессия

Наиболее нагляден (вплоть до возможности вычислений на калькуляторе) простейший случай модели (1), когда компоненты отклика описываются моделью

$$y_i = \theta_0 + \theta_1 g_1(t_i) + \varepsilon_i, \qquad i = 1, 2, \ldots, n. \tag{19}$$

Если допустить H': $\theta_1 = 0$, то регрессии y по t нет, а измерения y_i носят чисто случайный характер.

Предполагается выполнение требований классической модели. Посмотрим, к чему же приводит МНК для модели (19), и как проводится регрессионный анализ в случае нормальных ошибок. Помимо конкретизации общих результатов п. 2, мы обнаружим и специфические особенности для рассматриваемой ситуации. Чтобы обеспечить простоту и "красивость" формул, в дальнейшем мы будем предполагать, что регрессором является сам аргумент t, т.е. что $g_1(t) \equiv t$. Тогда в модели (19) всего $k = 2$ неизвестных параметра: свободный член θ_0 и коэффициент наклона θ_1.

Итак, в матричных обозначениях (19) записывается в виде (1′) с матрицей данных

$$\boldsymbol{X} = (\vec{\boldsymbol{1}}_{\boldsymbol{n}} : \vec{\boldsymbol{t}}) ,$$

где $\vec{\boldsymbol{1}}_{\boldsymbol{n}}$ – n-мерный вектор из единиц, $\vec{\boldsymbol{t}} = (t_1, \ldots, t_n)^T$. Тогда имеем

$$(\boldsymbol{X}^T\boldsymbol{X}) = \begin{pmatrix} n & n\bar{t} \\ n\bar{t} & \sum t_i^2 \end{pmatrix} ,$$

где выборочное среднее $\bar{t} = \frac{1}{n}\sum_{i=1}^{n} t_i = \frac{1}{n}\vec{\boldsymbol{1}}_{\boldsymbol{n}}^{\boldsymbol{T}}\vec{\boldsymbol{t}}$. Отсюда найдём

$$(\boldsymbol{X}^T\boldsymbol{X})^{-1} = \frac{1}{\sum (t_i - \bar{t})^2} \begin{bmatrix} \frac{1}{n}\sum t_i^2 & -\bar{t} \\ -\bar{t} & 1 \end{bmatrix} ,$$

а

$$\boldsymbol{X}^T\ \vec{\boldsymbol{y}}\ = \begin{bmatrix} \sum y_i \\ \sum t_i\, y_i \end{bmatrix}\ .$$

Следовательно, компоненты НК-оценки (3) равны

$$\hat{\theta}_0\ = \bar{y}\ -\ \hat{\theta}_1\ \bar{t}\ , \qquad (20а)$$

$$\hat{\theta}_1\ = \frac{\sum y_i\,(t_i - \bar{t})}{\sum (t_i - \bar{t})^2}\ = \frac{\sum (y_i - \bar{y})\ (t_i - \bar{t})}{\sum (t_i - \bar{t})^2}\ . \qquad (20б)$$

Подставляя в (6) $\hat{\theta}_0$ и $\hat{\theta}_1$, получим прогноз в точке t_i как

$$\hat{y}_i\ = \hat{\theta}_0\ +\ \hat{\theta}_1\ t_i\ = \bar{y}\ +\ \hat{\theta}_1\ (t_i - \bar{t}\ )\,. \qquad (21)$$

Конечно, вместо t_i в (21) можно рассматривать прогноз в произвольной точке t_*:

$$\hat{y}_*\ = \bar{y}\ +\ \hat{\theta}_1\ (t_* - \bar{t}\ )\,.$$

В частности, положив $t_*\ =\ \bar{t}$, получим, что предсказание проходит через точку $(\bar{t}\ ,\ \bar{y}\ )$.

С учётом (21) имеем

$$RSS = \|\ \vec{\boldsymbol{y}} - \hat{\boldsymbol{y}}\ \|^2\ = \sum_{i=1}^{n}(y_i - \bar{y})^2\ +\ \hat{\theta}_1^{\ 2} \sum_{i=1}^{n}(t_i - \bar{t})^2\ -2\ \hat{\theta}_1\ \sum_{i=1}^{n}(y_i - \bar{y})(t_i - \bar{t})$$

$$= \sum_{i=1}^{n}(y_i - \bar{y})^2\ -\ \hat{\theta}_1^{\ 2}\ \sum_{i=1}^{n}(t_i - \bar{t})^2\ , \qquad (22)$$

откуда

$$\hat{\sigma}^2\ = \frac{RSS}{n-2}\ = \frac{1}{n-2}\sum\left[(y_i - \bar{y})^2\ -\ \hat{\theta}_1^{\,2}(t_i - \bar{t})^2\right]\ . \qquad (23)$$

С учётом (21) формула (22) записывается в виде

$$RSS\ = \sum_{i=1}^{n}(y_i - \bar{y})^2\ -\ \sum_{i=1}^{n}(\hat{y}_i - \bar{y})^2\ , \qquad (22')$$

и тогда

$$SS_T\ = \sum_{i=1}^{n}(y_i - \bar{y})^2\ = \sum_{i=1}^{n}(\hat{y}_i - \bar{y})^2\ +\ \sum_{i=1}^{n}(y_i - \hat{y}_i)^2\ . \qquad (24)$$

Таким образом, полная сумма квадратов отклонений (Sum of Squares, Total), скорректированная на среднее, SS_T, представляется в виде двух слагаемых: первый член в правой части (24) – это сумма квадратов, обусловленная регрессией, а второе – обычная RSS (остаточная сумма квадратов).

Введём величину

$$R^2\ = \frac{\sum(\hat{y}_i - \bar{y})^2}{\sum(y_i - \bar{y})^2}\ = \hat{\theta}_1^{\ 2}\ \frac{\sum(t_i - \bar{t})^2}{\sum(y_i - \bar{y})^2}\ = \frac{\left[\sum(y_i - \bar{y})(t_i - \bar{t})\right]^2}{\sum(y_i - \bar{y})^2 \sum(t_i - \bar{t})^2}\ . \qquad (25)$$

Это не что иное, как квадрат выборочного коэффициента корреляции (*коэффициент детерминации* – частный случай квадрата множественной корреляции <см. [3], гл. 6>).

Формулу (22′) можно переписать с учётом (25) как

$$RSS\ = (1 - R^2)\ \sum_{i=1}^{n}(y_i - \bar{y})^2\ = (1 - R^2)\ SS_T. \qquad (22'')$$

Чем ближе R^2 к 1, тем меньше *RSS* . Но интуитивно ясно, что чем меньше доля *RSS* в SS_T, тем лучше регрессионная модель аппроксимирует данные.

С другой стороны, если случайные величины ξ и η линейно зависимы: $\eta = c\,\xi + d$, то из теории вероятностей вам известно, что квадрат коэффициента корреляции $\rho^2 = \mathrm{corr}^2\{\xi, \eta\} = 1$. Таким образом, R^2 является статистическим "двойником" ρ^2, характеризующим близость к линейной зависимости. Если же $R^2 \approx 0$, то по (25) $\hat{\theta}_1^{\,2} \approx 0$, так что y_i не зависят от t_i , чисто случайны. В случае нормальной регрессии можно придать точный смысл словам: " R^2 близок к нулю" или " R^2 близок к единице".

Разложение (24) SS_T на две суммы квадратов играет фундаментальную роль в нормальной регрессии, а также в дисперсионном анализе (**ANOVA**). С каждой из сумм в (24) можно связать целое число – число степеней свободы (**d.f.**) [1]. Это число показывает, как много независимых элементов информации, получающихся из n независимых чисел y_1 , y_2 , …, y_n требуется для образования данной суммы квадратов. Например, для SS_T требуется $(n - 1)$ независимый элемент (т.к. из чисел $y_1 - \bar{y}$, $y_2 - \bar{y}$, …, $y_n - \bar{y}$ независимы только $n - 1$: $\sum_1^n y_i = n\,\bar{y}$). Мы можем вычислить *SS* , обусловленную регрессией, используя единственную функцию от y_1 , y_2 , …, y_n , а именно, $\hat{\theta}_1$, в силу (21): для $i = 1, 2, \ldots, n$

$$\hat{y}_i - \bar{y} = \hat{\theta}_1\ (t_i - \bar{t}\ ),$$

где $\hat{\theta}_1 = \hat{\theta}_1\ (y_1, y_2, \ldots, y_n\ )$ определена в (20б). Таким образом,

$$\sum_{i=1}^{n} (\hat{y}_i - \bar{y})^2$$

имеет одну d.f. Можно показать [2, 3], что *RSS* имеет $n - 2$ d.f.

Разделив соответствующую *SS* на её d.f., получаем средний квадрат, MS. Так, для *RSS* величина MS равна $RSS/(n-2) = \hat{\sigma}^2$ (ср. с (9) и (23))

Эти результаты можно представить в форме таблицы ANOVA (дисперсионного анализа) [1]. При проведении регрессионного анализа средством "Анализ данных" Excel такая таблица ANOVA фигурирует среди результатов, и необходимо с ней разобраться. Основное разложение следующее:

ANOVA

Источник вериации	**Число d.f.**	**Суммы квадратов SS**	**Средние квадраты MS**
Обусловленный регрессией	1	$SS_{reg} = \sum_{i=1}^{n} (\hat{y}_i - \bar{y})^2$	$MS_{reg} = SS_{reg}$
Остаток относительно регрессии	$n - 2$	$RSS = \sum_{i=1}^{n} (y_i - \hat{y}_i)^2$	$\hat{\sigma}^2 = \frac{RSS}{n-2}$
Полный, скорректированный на среднее	$n - 1$	$SS_T = \sum_{i=1}^{n} (y_i - \bar{y})^2$	$MS_T = \frac{SS_T}{n-1}$

В случае повторных наблюдений в некоторых точках t_i возможны и более детальные разложения ANOVA (см. в конце этого раздела).

При рассмотрении нормальной регрессии можно вычислить интервальные оценки для параметров θ_0 и θ_1 согласно (16). Так, если для

заданного α $t_{n-2,\,1-\alpha/2}$ – $(1-\alpha/2)$-квантиль распределения Стьюдента с $n-2$ d.f., то доверительный интервал для θ_0 есть

$$\hat{\theta}_0 \pm t_{n-2,\,1-\alpha/2} \left\{ \frac{\sum t_i^2}{n \sum (t_i - \bar{t})^2} \right\}^{1/2} \hat{\sigma} \quad , \tag{26}$$

а для θ_1 получим доверительный интервал

$$\hat{\theta}_1 \pm t_{n-2,\,1-\alpha/2} \left\{ \frac{\hat{\sigma}^2}{\sum (t_i - \bar{t})^2} \right\}^{1/2} \quad . \tag{27}$$

В частности, можно использовать доверительный интервал (27) для проверки значимости регрессии – гипотезы H': $\theta_1 = 0$; если точка 0 является внутренней точкой для интервала (27), то H' принимается, в противном случае H' отвергается.

Иначе можно проверять H' по значению F-отношения (эта величина и критическое значение для неё тоже подсчитываются средствами "Анализ данных" Excel): если вычисленное по (18) значение F превышает критическое $F_{1,\,n-2,\,1-\alpha}$, то H' отвергается, и разумно считать $\theta_1 \neq 0$. Так как для H' число ограничений $q = 1$, вектор $\vec{c} = 0$, $\boldsymbol{A} = (0, 1) = \vec{e}_2^{\,T}$, то $\boldsymbol{B} = \vec{e}_2^{\,T} (\boldsymbol{X}^T \boldsymbol{X})^{-1} \vec{e}_2 = (X^T X)^{-1}_{22} = \dfrac{1}{\sum (t_i - \bar{t})^2}$, так что

$$F = \frac{\hat{\theta}_1^2}{\hat{\sigma}^2} \{(X^T X)^{-1}_{22}\}^{-1} = \frac{\hat{\theta}_1^2}{\hat{\sigma}^2} \sum_{i=1}^{n} (t_i - \bar{t})^2 = \frac{R^2 (n-2)}{1-R^2}$$

вследствие (22"). Обычная t-статистика для проверки H': $\theta_1 = 0$ (независимости $\vec{y}$ и $\vec{t}$) имеет вид

$$T = R \sqrt{\frac{n-2}{1-R^2}} \quad , \tag{28}$$

так что $F = T^2$.

Для рассмотрения двух последних обсуждаемых вопросов этого раздела нам понадобится

Лемма о переносе ошибок. Пусть с.в. $\vec{\xi} \in R^m$, $E\{\vec{\xi}\} = \vec{\alpha}$, $\mathrm{cov}\{\vec{\xi}, \vec{\xi}\} = \boldsymbol{C}$. Рассмотрим l-мерную с.в. $\vec{\eta} = \boldsymbol{L}\,\vec{\xi}$, матрица $\boldsymbol{L} \in R^{l \times m}$. Тогда

(а) $E\{\vec{\eta}\} = \boldsymbol{L}\vec{\alpha}$;

(б) $\mathrm{cov}\{\vec{\eta}, \vec{\eta}\} = \boldsymbol{L}\,\boldsymbol{C}\,\boldsymbol{L}^T$.

Доказательство можно найти, например, в [3].

Теперь вернёмся к обсуждаемой теме. Так как $\hat{y} = \boldsymbol{X}\,\hat{\boldsymbol{\theta}}$ и $\hat{\boldsymbol{\theta}}$ – несмещённая оценка , то $E\{\hat{y}\} = \boldsymbol{X}\,\vec{\boldsymbol{\theta}} = E\{\vec{y}\}$, и вектор регрессионных остатков $\vec{e} = \vec{y} - \hat{y}$ имеет $E\{\vec{e}\} = \vec{0}$, что используется при анализе остатков.

В модели линейной одномерной регрессии прогноз $\hat{y}_*$ может быть записан в матричной форме как

$$\hat{y}_* = (1, t_*) \begin{pmatrix} \hat{\theta}_0 \\ \hat{\theta}_1 \end{pmatrix} = \vec{X}_*^{\,T} \hat{\boldsymbol{\theta}} \quad , \tag{29}$$

с вектором-строкой $\vec{X}_*^{\,T} = (1, t_*)$. Согласно утверждению (а) леммы,

$$E\{\hat{y}_*\} = \vec{X}_*^{\,T}\,\vec{\theta}\ , \qquad (30)$$

а по утверждению (б) дисперсия $\hat{y}_*$ равна

$$\sigma^2 v_* = Var\{\hat{y}_*\} = \sigma^2\ \vec{X}_*^{\,T}\,(\boldsymbol{X}^T\boldsymbol{X})^{-1}\,\vec{X}_* = \sigma^2\ \frac{\left\{\sum t_i^2 - 2t_*\, n\bar{t} + nt_*^2\right\}}{n\sum (t_i - \bar{t})^2}$$

$$= \sigma^2\ \frac{\left\{\sum t_i^2 - n\bar{t}^{\,2} + n(t_* - \bar{t})^2\right\}}{n\sum (t_i - \bar{t})^2} = \sigma^2\ \left(\frac{1}{n} + \frac{(t_* - \bar{t})^2}{\sum (t_i - \bar{t})^2}\right)\ . \qquad (31)$$

Как функция t_*, величина v_* минимальна при $t_* = \bar{t}$, и возрастает с удалением от $\bar{t}$. При практическом вычислении $Var\{\hat{y}_*\}$ величина σ^2 в (31) заменяется её оценкой $\hat{\sigma}^2$, и тогда центральный $100*(1-\alpha)$-процентный доверительный интервал для математического ожидания прогноза в точке t_* – величины в формуле (30) – имеет вид [1, 2]

$$\hat{y}_* - t_{n-2,\,1-\alpha/2}\ \hat{\sigma}\ \sqrt{v_*}\ \le E\{\hat{y}_*\} \le\ \hat{y}_* + t_{n-2,\,1-\alpha/2}\ \hat{\sigma}\ \sqrt{v_*}\ . \qquad (32)$$

Могут быть построены доверительные интервалы для предсказания самого отклика в точке t_* [1, 2].

Осталось рассмотреть ещё вопрос о корректности рассматриваемой регрессионной модели. В общем случае этот вопрос решается на основе анализа графиков регрессионных остатков, о чём подробно рассказано в [1, 2]. Но если имеются неоднократные измерения в одной и той же точке, то можно использовать эту дополнительную информацию для проведения статистического анализа по формальным критериям. Эта тема подробно рассматривается в [1, .п. 1.5].

Рассмотрим остатки $e_i = y_i - \hat{y}_i$ при $t = t_i$. Остатки содержат всю мыслимую информацию относительно того, почему построенная модель недостаточно правильно объясняет наблюдаемый разброс значений отклика $\vec{y}$. Пусть $\eta_i = E\{y_i\}$ – математическое ожидание (м.о.) для "истинной" модели при $t = t_i$. Тогда мы можем записать

$$e_i = y_i - \hat{y}_i = (y_i - \hat{y}_i) - E\{y_i - \hat{y}_i\} + E\{y_i - \hat{y}_i\}$$

$$= \{(y_i - \hat{y}_i) - (\eta_i - E\{\hat{y}_i\})\} + (\eta_i - E\{\hat{y}_i\}) = q_i + B_i\ , \qquad (33)$$

где

$$q_i = \{(y_i - \hat{y}_i) - (\eta_i - E\{\hat{y}_i\})\}\ ,\ B_i = \eta_i - E\{\hat{y}_i\}\ . \qquad (34)$$

Величина B_i в (34) – ошибка смещения при $t = t_i$. Если модель верна, то $E\{\hat{y}_i\} = E\{y_i\} = \eta_i$ и $B_i = 0$; если нет, то $B_i \neq 0$, и его значение зависит от "истинной" модели и значения t_i. Нетрудно видеть, что q_i – с.в. с нулевым математическим ожиданием, $E\{q_i\} = 0$, и это верно <u>независимо</u> от того, будет ли модель правильна.

Можно показать, что q_i коррелированы, и величина $q_1^{\,2} + q_2^{\,2} + \ldots + q_n^{\,2}$ имеет м.о. $(n-2)\,\sigma^2$. Отсюда следует, что остаточная сумма квадратов

$$\frac{1}{n-2}\sum_{i=1}^{n}(y_i - \hat{y}_i)^2$$

имеет м.о. σ^2, если постулированная модель корректна, и $\sigma^2 + \frac{1}{n-2}\sum B_i^2$, если модель не корректна.

Если модель не корректна ($B_i \neq 0$), то остатки (33) содержат оба компонента: случайный (q_i) и систематический (B_i). Мы можем рассматривать их соответственно как случайную ошибку разброса и систематическую ошибку смещения. Тогда возможны две ситуации, в которых оценивается вклад систематической ошибки смещения в *RSS* : (а) существует априорная оценка σ^2 (скажем, по ранее выполненным опытам); тогда можно проверить, значимо ли $RSS > \sigma^2$. В случае (б) априорной оценки σ^2 нет, но измерения y повторялись (два раза или более) при одинаковых значениях t; тогда можно построить апостериорную оценку σ^2. Про такую оценку говорят, что она представляет "чистую" ошибку, потому что если t одинаково для двух наблюдений, то только случайные вариации могут влиять на результаты и создавать разброс между ними.

Будем обозначать повторные наблюдения при одном и том же t дополнительным верхним индексом, так что если наблюдения были сделаны в m различных значениях t и к i-му из этих t относится n_i наблюдений, то

$y_1^{(1)}, y_1^{(2)}, \ldots, y_1^{(n_1)}$ обозначает n_1 повторных наблюдений при t_1;

$y_2^{(1)}, y_2^{(2)}, \ldots, y_2^{(n_2)}$ обозначает n_2 повторных наблюдений при t_2;

.....

$y_i^{(1)}, y_i^{(2)}, \ldots, y_i^{(n_i)}$ обозначает n_i повторных наблюдений при t_i;

.....

$y_m^{(1)}, y_m^{(2)}, \ldots, y_m^{(n_m)}$ обозначает n_m повторных наблюдений при t_m

Замечание. Повторные наблюдения могут проводиться не во всех точках $t_1, \ldots, t_m$. Если t_l – такая точка, для которой сделано единственное наблюдение, то верхний индекс будет опускаться, а в нижеследующих формулах подразумевается, что он принимает единственное значение – единица.

Всего получается

$$n = \sum_{i=1}^{m} \sum_{u=1}^{n_i} 1 = \sum_{i=1}^{m} n_i$$

наблюдений. Вклад суммы квадратов, связанной с "чистой" ошибкой для n_1 наблюдений при t_1, равен внутренней сумме квадратов $y_1^{(u)}$ относительно их среднего $\bar{y}_1$, т.е.

$$\sum_{u=1}^{n_1} (y_1^{(u)} - \bar{y}_1)^2 = \sum_{u=1}^{n_1} (y_1^{(u)})^2 - n_1 \bar{y}_1^2 = \sum_{u=1}^{n_1} (y_1^{(u)})^2 - \frac{1}{n_1} \left(\sum y_1^{(u)} \right)^2 , \qquad (35)$$

и точно так же будет для $t_2, \ldots, t_m$.

Объединяя внутренние суммы квадратов для всех серий повторных опытов, получим общую сумму квадратов

$$\sum_{i=1}^{m} \sum_{u=1}^{n_i} (y_i^{(u)} - \bar{y}_i)^2 \qquad (36)$$

со степенями свободы

$$n_e = \sum_{i=1}^{m} (n_i - 1) = \sum_{i=1}^{m} n_i - m . \qquad (37)$$

Отсюда средний квадрат "чистых" ошибок равен

$$s_e^2 = \frac{\sum_{i=1}^{m}\sum_{u=1}^{n_i}(y_i^{(u)} - \bar{y}_i)^2}{\sum_{i=1}^{m} n_i - m}, \qquad (38)$$

и он служит оценкой σ^2 безотносительно к тому, корректна ли подобранная модель.

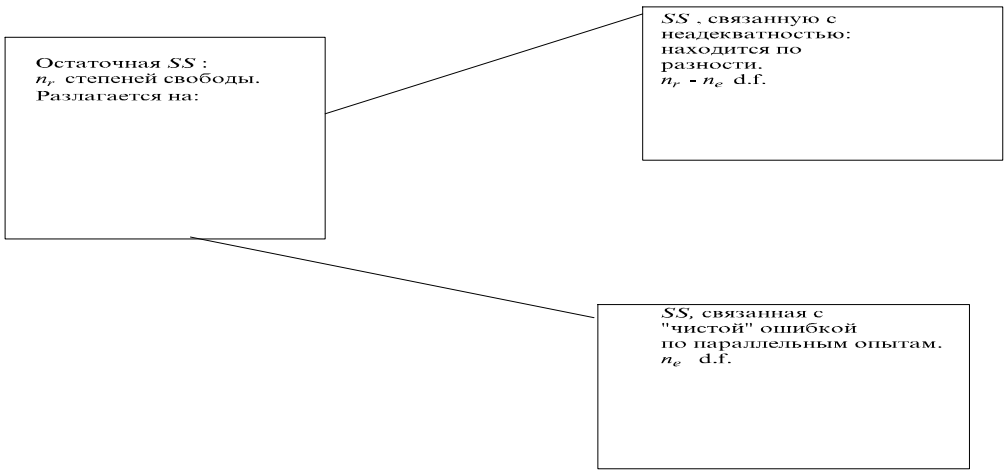
Остаток для u-го наблюдения при t_i можно записать в виде

$$y_i^{(u)} - \hat{y}_i = (y_i^{(u)} - \bar{y}_i) - (\hat{y}_i - \bar{y}_i) \qquad (39)$$

(все повторные точки при любом t_i имеют одно и то же предсказанное значение $\hat{y}_i$). Возведём в квадрат обе части (39), а затем просуммируем их по u и по i. Получим

$$\sum_{i=1}^{m}\sum_{u=1}^{n_i}(y_i^{(u)} - \hat{y}_i)^2 = \sum_{i=1}^{m}\sum_{u=1}^{n_i}(y_i^{(u)} - \bar{y}_i)^2 + \sum_{i=1}^{m} n_i (\hat{y}_i - \bar{y}_i)^2 \qquad (40)$$

(перекрёстные произведения исчезают при суммировании по u для каждого i). Слева в (40) – *RSS*. Двойная сумма в правой части – это сумма квадратов "чистых" ошибок. Последнюю (однократную) сумму называют суммой квадратов неадекватности. Отсюда следует, что сумму квадратов, обусловленную "чистой" ошибкой, и неадекватностью, можно представить в виде рисунка:



Обозначим сумму квадратов, связанную с неадекватностью, через SS_L ; поделив её на связанное с ней d.f. $n_L = (n_r - n_e)$, получим средний квадрат ошибки неадекватности MS_L. Обычный приём – это сравнение отношений

$$F = MS_L / s_e^2 \qquad (41)$$

со $100\cdot(1-\alpha)$ %-ной точкой F-распределения при $(n_r - n_e)$ и n_e d.f. Если это отношение:

1) значимо, то модель, по-видимому, неадекватна. Исследуя графики остатков, можно попытаться понять, когда и как встречается неадекватность (см. [1]);
2) незначимо; тогда нет оснований для сомнений в адекватности модели и что как средний квадрат, связанный с "чистой" ошибкой s_e^2, так и

средний квадрат, обусловленный неадекватностью, могут использоваться как оценки σ^2. Объединённая оценка σ^2 может быть получена из суммы квадратов, связанной с неадекватностью, и суммы квадратов, связанной с "чистой" ошибкой, путём объединения их в остаточную сумму квадратов и деления её на число d.f. n_r, что даёт

$$\hat{\sigma}^2 = (\text{"остаточная" } SS) / n_r . \qquad (42)$$

Подытожим все необходимые действия, когда наши данные содержат повторные наблюдения:

1. Подобрать модель, составить простую таблицу ANOVA с двумя входами: регрессией и остатком. Но для общей регрессии пока что **не** использовать *F*-критерий.
2. Вычислить *SS*, связанную с "чистой" ошибкой, и разложить *RSS*, как на рис.
3. Применить *F*-критерий для неадекватности (41). Если вычисленное значение не значимо, т.е. нет смысла сомневаться в адекватности модели, то перейти к пункту 5.
4. Значимая неадекватность. Прекратить анализ подобранной модели и искать пути улучшения модели методами анализа остатков. **Не применять** *F*-критерий для общей регрессии и **не** пытаться строить доверительные интервалы < ибо исходные предпосылки для этих процедур не верны>.
5. Неадекватность не значима. Снова объединить суммы квадратов для "чистых" ошибок и неадекватности в *RSS*. Использовать остаточный средний квадрат $\hat{\sigma}^2$ (42), применить *F*-критерий для общей регрессии, получить доверительные пределы для $E\{\hat{y}_*\}$, вычислить R^2 и т.д. А графики остатков всё-таки надо строить и надо исследовать их особенности [1, гл. 3].

4. Примеры анализа с помощью Excel

Рассмотрим, как вычислять НК-оценки и проводить регрессионный анализ средствами Excel. Как ни странно, но в "ширпотребовском" Excel есть очень мощное средство – "Анализ данных" в меню **Сервис**, которое позволяет не только найти НК-оценки, но и выполнить почти весь описанный в п. 3 регрессионный анализ (исключая исследование неадекватности при наличии повторных наблюдений), посмотреть критические значения для различных статистик, полюбоваться некоторыми графиками остатков, увидеть, каково предсказание… И всё это – двумя движениями мыши, чтобы указать диапазоны для независимой переменной и отклика – остальное сделает сама программа. А даже в таких "интеллектуальных" программах, как Maple или MATLAB, придётся потрудиться, чтобы написать программу регрессионного анализа (в этих средствах предусмотрено лишь вычисление НК-оценок, но даже оценку $\hat{\sigma}^2$ приходится вычислять самому), а потом ещё организовывать демонстрацию различных графиков: подгонки, остатков и т.п.

Работу в Excel я иллюстрирую на двух примерах, рассматриваемых в [1]. Такой выбор позволит вам сравнивать получаемые вами результаты с приведёнными в книге, задаваться вопросами для дальнейших исследований. Исходные данные для этих примеров записаны в файле demo_ls.xls.

Предполагается, что у вас есть некоторые навыки работы в Excel. Но так как возможности "Анализа данных" известны немногим, то возможно, что вам потребуется некоторая предварительная подготовка, чтобы им пользоваться на вашем домашнем или рабочем компьютере. Обычно это средство уже есть "в скрытом виде" при обычной установке Microsoft Office; если вы обнаружите, что указанные ниже действия не приводят к успеху, то целесообразно проконсультироваться со специалистами… а, может быть, установить Office заново.

Итак, на моём компьютере в меню **Сервис** последней опцией был "Анализ данных", но он был недоступен (отображался не чёрным, а серым цветом). Я узнал в службе Help, что нужно сделать, чтобы эта опция стала активной; для этого нужно иметь файл Analys32.xll в папке Excel \ Library \ Analysis. Оказалось, что соответствующий файл Analys32 уже был помещён в папку Excel \ Library \ Analysis при постановке Office. Проверьте, так ли это на вашем компьютере. Если так, то для активизации "Анализа данных" нужно выбрать в меню **Сервис** (**Tools** в английской версии) пункт "Настройки" и установить флаг ("галочку") на "Пакет анализа", щёлкнуть OK. При вторичном обращении к **Сервису** опция "Анализ данных" будет показана уже как активная (чёрным цветом).

Первый пример – "сквозной" пример книги Дрейпера и Смита, анализируемый в гл. 1 и далее. В выпарном аппарате на промышленном предприятии y – количество используемого пара в фунтах ежемесячно; в качестве регрессоров можно предложить несколько величин, в том числе t – среднемесячная температура воздуха (° F); см. также задачу 185 практикума. В качестве модели рассматривается линейная одномерная регрессия. В течение 25 месяцев было сделано 25 наблюдений:

t	35.3	29.7	30.8	58.8	61.4	71.3	74.4
y	10.98	11.13	12.51	8.40	9.27	8.73	6.36
t	76.7	70.7	57.5	46.4	28.9	28.1	39.1
y	8.50	7.82	9.14	8.24	12.19	11.88	9.57
t	46.8	48.5	59.3	70.0	70.0	74.5	72.1
y	10.94	9.58	10.09	8.11	6.83	8.88	7.68
t	58.1	44.6	33.4	28.6			
y	8.47	8.86	10.36	11.08			

Итак, работаем с первым примером (**Example 1**) файла demo_ls.xls. Данные расположены в столбцах A и B, в строках 3 : 27. Вызываем "Анализ данных", метод анализа – "Регрессия". После того, как откроется соответствующее диалоговое окно, нужно указать (набрав на клавиатуре или выделив мышью) Входной интервал Y < b3:b27 >, Входной интервал X (он соответствует переменной t в наших обозначениях) <a3:a27 >, остальные входные параметры оставить так, как их указывает программа (Уровень надёжности 95 % < это не что иное, как наша $\gamma = 1 - \alpha$, так что работаем со стандартным уровнем значимости α = 5 % >, флаг "константа – ноль" – в положении "off ").

В параметрах вывода нужно указать: radio-button "Новый рабочий лист" нажатой (стоит чёрная точка в центре), как по умолчанию < ваши результаты

будут выводиться в лист 17>; следует активизировать <"on", √ > следующие средства показа остатков:

Остатки

√ Остатки

Стьюдентизированные остатки

√ График остатков

√ График подбора

Остальные возможности вывода – Стьюдентизированные остатки, График нормальной вероятности – на ваше усмотрение.

После того, так вы установите все указанные выше параметры ввода и вывода, в Листе 17 вы увидите следующие таблицы итогов (обозначения "Таблица 1", "Таблица 2" и т.д., и комментарии к результатам – мои, всё остальное - работа Excel). Практический совет: форма выдачи итогов, используемая Excel по умолчанию, довольно неудобна. Поэтому: 1) максимально уменьшите размер символов (на моём компьютере - 8 пунктов), чтобы быть в состоянии прочитать то, что распечатано ниже; 2) измените ширины столбцов A, B, C, D и т.д. с помощью мыши, двигая эти заголовки A, B, C, … , как обычно в программах приложений Windows – иначе вы не сможете прочитать заголовки: они слишком узкие.

Итак, вывод результатов: Таблица 1

ВЫВОД ИТОГОВ

Регрессионная статистика	
Множественный R	0,84524406
R-квадрат	0,714437521
Нормированный R-квадрат	0,702021761
Стандартная ошибка	0,890124516
Наблюдения	25

Мой комментарий: почти всё в ней понятно. Вычислены R^2 (формула (25)) и корень из него ("Множественный R), в [1] поясняется, что такое "Нормированный R-квадрат", но авторы считают его мало информативным и не рекомендуют использовать. Далее приведена "стандартная ошибка" – это корень из оценки $\hat{\sigma}^2$ общей дисперсии (формула (23)); наблюдений у нас 25.

Таблица 2 – результаты дисперсионного анализа

ANOVA

	df	*SS*	*MS*	*F*	*Значимость F*
Регрессия	1	45,592402	45,592402	57,542794	1,055E-07
Остаток	23	18,223398	0,7923217		
Итого	24	63,8158			

Она в основном соответствует основному разложению ANOVA §3 этого руководства, только добавлены два столбца: "*F"* и *"*Значимость *F"*. В основном заголовки совпадают с заголовками моей таблицы (но, естественно, названия даны в краткой форме: например, "Регрессия" ≡ "Обусловленный регрессией"; "Итого" ≡ "Полный, скорректированный на среднее").

Здесь тоже всё понятно: так как у нас всего $n = 25$ наблюдений, то в столбце 2 фигурируют d.f.: для "Итого" $n - 1 = 24$, для "Остаток" $n - 2 = 23$, а

число d.f. в "Регрессии" всегда равно 1. В столбце 3 приведены суммы квадратов в полном соответствии с (24). Видно, что *RSS* не слишком мала (об этом свидетельствует и довольно скромное значение $R^2 = 0.71$). На мой взгляд, это свидетельства того, что одного регрессора может быть недостаточно для объяснения наблюдаемых данных, и потребуются более сложные модели, чем линейная одномерная регрессия (авторы так и сделали позже в гл. 4, введя две предикторные переменные). В 4-ом столбце даны “Средние квадраты ошибок”, из которых наиболее интересно число в строке "Остаток" – это $\widehat{\sigma}^2$ (возведите в квадрат "Стандартную ошибку" таблицы 1 и убедитесь, что эта величина равна 0,792321654 с точностью до ошибок округлений). Вычисленное в столбце 5 значение $F = 57,54279428$ очень велико и превышает критическое значение $F_{1\ ,\ 23\ ,\ 0.95} = 4.279343102$ (я вычислил его с помощью функции Excel FРАСПОБР). Таким образом, гипотеза H' : $\theta_1 = 0$ отклоняется. Разработчики "Анализа данных" пошли по другому пути, чем я. Они не приводят критического значения для *F*-распределения, но вычисляют вероятность того, что при условии, что H': $\theta_1 = 0$ верна, может получиться значение $F \geq 57,54279428$. Эта вероятность исчезающе мала: "Значимость $F'' = 1,054995 \cdot 10^{-7}$, так что регрессия значима ($\theta_1 \neq 0$).

В следующей Таблице 3 даны оценки параметров с указанием их стандартных ошибок, доверительные границы для них, и ещё некоторые результаты. Об обозначениях: Y-пересечение – это "свободный член θ_0 в моей

	Коэффициенты	*Стандартная ошибка*	*t-статистика*	*P-Значение*	*Нижние 95%*	*Верхние 95%*
Y-пересечение	13,62298927	0,5814635	23,428795	1,497E-17	12,420142	14,825837
Переменная X 1	-0,079828693	0,0105236	-7,5856967	1,055E-07	-0,1015983	-0,058059

терминологии, а Переменная X 1 – это θ_1 , коэффициент наклона. Приведены их НК-оценки ("Коэффициенты"), стандартные ошибки для них, левая ("Нижние 95 %") и правая ("Верхние 95 %") границы доверительного интервала для параметров (формулы (26) и (27)). Что такое "*t*-статистика" и *"P*-значение", для меня пока что не полностью ясно. Скорее всего, проверяется гипотеза, что значение параметра $\theta_0 = 0$ (или, соответственно, $\theta_1 = 0$) и вероятность того. что вычисленное или даже большее (по модулю) значение *t*-статистики возможно, если проверяемая гипотеза верна. По крайней мере, в [1] приведено значение $t = -7.60$ для параметра θ_1 (проверяется гипотеза H': $\theta_1 = 0$). Видно, что обе вероятности микроскопически малы, так что гипотеза о нулевых значениях параметров отклоняется. Но мы, как было показано при обсуждении доверительного интервала (27), имеем другой способ проверки гипотезы о равенстве 0 соответствующего параметра: если точка 0 является внутренней для доверительного интервала, то гипотеза принимается, в противном случае отклоняется. Согласно этому правилу, оба параметра отличны от 0. Для меня осталось неясным, зачем столбцы "Нижние 95 %" и "Верхние 95 %" напечатаны многократно, но это одна из многочисленных загадок Excel. К сожалению, доверительный интервал для м.о. прогноза среди итогов регрессионного анализа в Excel не вычисляется, и если в задаче будет такой вопрос, придётся это сделать вручную.

Наконец, в Таблице 4 для каждого наблюдения 1, 2, …, 25 распечатаны предсказание и (регрессионный) остаток.

ВЫВОД ОСТАТКА

Наблюдение	*Предсказанное Y*	*Остатки*
1	10,80503639	0,1749636
2	11,25207708	-0,1220771
3	11,16426551	1,3457345
4	8,929062101	-0,5290621
5	8,721507499	0,5484925
6	7,931203435	0,7987966
7	7,683734486	-1,3237345
8	7,500128491	0,9998715
9	7,979100651	-0,1591007
10	9,032839403	0,1071606
11	9,918937899	-1,6789379
12	11,31594003	0,87406
13	11,37980299	0,500197
14	10,50168736	-0,9316874
15	9,887006421	1,0529936
16	9,751297643	-0,1712976
17	8,889147755	1,2008522
18	8,034980736	0,0750193
19	8,034980736	-1,2049807
20	7,675751616	1,2042484
21	7,86734048	-0,1873405
22	8,984942187	-0,5149422
23	10,06262955	-1,2026295
24	10,95671091	-0,5967109
25	11,33988864	-0,2598886

Но нагляднее, чем эти сухие цифры, график подбора (исходные данные и прогноз – на одном графике) и график остатков (в зависимости от значений регрессора t). Конечно, это далеко не исчерпывающее исследование остатков, но имея числовые данные по остаткам в Таблице 4, в Excel нетрудно построить и другие полезные при анализе остатков графики (см. [1, 2]).

Пример2. Если в примере 1 рассматривались данные для вполне реального практического примера, то анализ адекватности при наличии повторных наблюдений мы, вслед за авторами [1], проведём для специально построенного примера с 24 наблюдениями.

Исходные данные представлены в форме следующей таблицы:

? наблюдения	y	t	? наблюдения	y	t	? наблюдения	y	t
1	2.3	1.3	9	1.7	3.7	17	3.5	5.3
2	1.8	1.3	10	2.8	4.0	18	2.8	5.3
3	2.8	2.0	11	2.8	4.0	19	2.1	5.3
4	1.5	2.0	12	2.2	4.0	20	3.4	5.7
5	2.2	2.7	13	5.4	4.7	21	3.2	6.0
6	3.8	3.3	14	3.2	4.7	22	3.0	6.0
7	1.8	3.3	15	1.9	4.7	23	3.0	6.3
8	3.7	3.7	16	1.8	5.0	24	5.9	6.7

Требуется подогнать данные моделью линейной одномерной регрессии и решить, адекватна ли ситуации эта модель.

Прежде всего, перейдём из Листа 17 в Лист 1, где находятся исходные данные. Расположение данных такое же, как в вышеприведённой таблице: в столбце A – номера наблюдений, отклик – в столбце B, а в столбце C – значения регрессора *t*. Данные занимают строки 31 : 54.

Точно так же, как в примере 1, задаём входные параметры и желаемые выходные параметры. Вызываем "Регрессию" в "Анализе данных" и получаем следующие результаты:

ВЫВОД ИТОГОВ

Регрессионная статистика	
Множественный R	0,479410745
R-квадрат	0,229834663
Нормированный R-квадрат	0,194827147
Стандартная ошибка	0,981503172
Наблюдения	24

ANOVA

	df	*SS*	*MS*	*F*	*Значимость F*
Регрессия	1	6,324666857	6,324666857	6,56529492	0,017765991
Остаток	22	21,19366648	0,963348476		
Итого	23	27,51833333			

	Коэффициенты	*Стандартная ошибка*	*t-статистика*	*P-Значение*	*Нижние 95%*	*Верхние 95%*
Y-пересечение	1,436395499	0,590007206	2,434538907	0,02348062	0,212794131	2,659996868
Переменная X 1	0,337886218	0,131869196	2,562283146	0,01776599	0,064405951	0,611366485

Эти результаты оставляют двойственное впечатление: с одной стороны, вычисленное *F*-отношение является значимым (критическое значение $F_{1\ ,\ 22\ ,\ 0.95} = 4.30$, а у нас $F =$ 6,56529492), так что нет оснований считать $\theta_1 = 0$, и регрессия *y* по *t* значима, а не имеет места чисто случайный разброс "игреков". С другой стороны, величина R^2 намного меньше 1, да и доля "Остатка" в "Итого" подозрительно велика, так что есть сомнения, что рассматриваемая модель адекватна данным. Изучим этот вопрос, используя имеющиеся повторные измерения. Если вы посмотрите на "График подбора", то увидите, что точки наблюдений далеки от прогноза именно там, где есть повторные измерения (как правило, они разные, и невозможно согласовать прогноз в данной точке с различающимися наблюдениями). Естественно оценить, какова доля случайного разброса, а какая часть приходится на (возможную) неадекватность модели. Для анализа вам придётся потрудиться вручную, чтобы вычислить выборочные средние и разбросы в точках повторных измерений.

Можно выполнить все требуемые расчёты, оставаясь в Лист 1 Excel. Имеем (R_t – разброс данных в точке t) :

$t = 1.3$ $\quad \bar{y}_{1.3} = 2.05$ $\quad R_{1.3} = 0.125$ $\quad 1$ *d.f. :*

$t = 2.0$ $\quad \bar{y}_{2.0} = 2.15$ $\quad R_{2.0} = 0.845$ $\quad 1$ *d.f. :*

$t = 3.3$ $\quad \bar{y}_{3.3} = 2.80$ $\quad R_{3.3} = 2.0$ $\quad 1$ *d.f. :*

$t = 3.7$ $\quad \bar{y}_{3.7} = 2.70$ $\quad R_{3.7} = 2.0$ $\quad 1$ *d.f. :*

$t = 4.0$	$\bar{y}_{4.0} = 2.60$	$R_{4.0} = 0.24$	2 *d.f. :*
$t = 4.7$	$\bar{y}_{4.7} = 3.50$	$R_{4.7} = 6.26$	2 *d.f. :*
$t = 5.3$	$\bar{y}_{5.3} = 2.80$	$R_{5.3} = 0.98$	2 *d.f. :*
$t = 6.0$	$\bar{y}_{6.0} = 3.10$	$R_{6.0} = 0.02$	1 *d.f. .*

Просуммировав R_t и степени свободы, получим сумму квадратов для "чистых" $SS_e = 12.47$ и эта величина имеет $n_e = 11$ *d.f. ;* средний квадрат "чистых" ошибок равен $s_e^2 = 1.133636$. Отсюда сумма квадратов, связанная с неадекватностью, получается $21{,}19366648 - SS_e = 8.723666$, и ей соответствует $MS_L = 0.793061$. Вычисляя F-отношение (41)

$$F = MS_L / s_e^2 \quad ,$$

Видим, что оно явно незначимо: $F = 0.699572 < 1$, а для значимости, во всяком случае, необходимо, чтобы $F > 1$. Следовательно, предложенная модель оказывалась адекватной.

ЛИТЕРАТУРА

Дрейпер Н., Смит Г. *Прикладной регрессионный анализ*. Изд. второе. Книга 1. М.: Финансы и Статистика. 1986

Себер Дж. *Линейный регрессионный анализ*. М.: Мир. 1980

Уфимцев М.В. *Методы многомерного статистического анализа* . М.: Издательский отдел ф-та ВМиК МГУ. 1997