МИНИСТЕРСТВО ОБРАЗОВАНИЯ И НАУКИ
РОССИЙСКОЙ ФЕДЕРАЦИИ

КУРГАНСКИЙ ГОСУДАРСТВЕННЫЙ УНИВЕРСИТЕТ

Кафедра «Автомобильный транспорт и автосервис»

ОБРАБОТКА РЕЗУЛЬТАТОВ ЭКСПЕРИМЕНТА С ПОМОЩЬЮ MICROSOFT EXCEL

ЧАСТЬ 2

Лабораторный практикум
по дисциплине «Компьютерные технологии в науке и образовании» для студентов квалификации (степени) Магистр направления 190500 (552100)

Курган 2011

Кафедра: «Автомобильный транспорт и автосервис»
Дисциплина: «Компьютерные технологии в науке и образовании» специализированной подготовки направления 190600 (552100) «Эксплуатация транспортных средств» (квалификация выпускника (степень) магистр)
Составил: канд. техн. наук, доцент Д.И. Дик

Утвержден на заседании кафедры «6» мая 2011 г.

Рекомендован методическим советом университета «2» июня 2011 г.

СОДЕРЖАНИЕ

1 ВВЕДЕНИЕ В РЕГРЕССИОННЫЙ АНАЛИЗ

В регрессионном анализе рассматривается связь между одной переменной, называемой зависимой переменной и несколькими другими, называемыми независимыми переменными. Эта связь представляется с помощью математической модели, т.е. уравнения, которое связывает зависимую переменную с независимыми с учетом множества соответствующих предположений. Независимые переменные связаны с зависимой посредством функции регрессии, зависящей также от набора неизвестных параметров. Если функция линейна относительно параметров (но необязательно линейна относительно независимых переменных), то говорят о линейной модели регрессии. В противном случае модель называется нелинейной. В каждом из этих случаев говорят о регрессии зависимой переменной по независимым переменным.

Статистическими проблемами регрессионного анализа являются:

а) получение наилучших точечных и интервальных оценок неизвестных параметров регрессии;

б) проверка гипотез относительно этих параметров;

в) проверка адекватности предполагаемой модели;

г) проверка множества соответствующих предположений.

Выбор подходящей модели основывается скорее не на статистических доводах, а на основе учета физических факторов. Далее будут обсуждаться некоторые аналитические средства, полезные при определении зависимости между переменными.

Регрессионный анализ используется по двум причинам. Во-первых, потому, что описание зависимости между переменными помогает установить наличие возможной причинной связи. Во-вторых, для получения предиктора для зависимой переменной, так как уравнение регрессии позволяет предсказывать значения зависимой переменной по значениям независимых переменных. Эта возможность особенно важна в тех случаях, когда прямые измерения зависимой переменной затруднены или дорого стоят.

1.1 Простая линейная регрессия

Рассмотрим ситуацию, когда две переменные связаны линейным соотношением.

Пусть Y — зависимая, а X — независимая переменные. Предположим, что имеется выборка парных наблюдений (x_1, y_1), (x_2, y_2), …, (x_n, y_n) из некоторой популяции W.

Первый способ получения выборки из популяции состоит в том, что значения X фиксируются, скажем, $X = x_1$, …, $X = x_n$, так что для $X = x_i$, мы имеем подпопуляцию W_i из W, содержащую все индивидуумы, для которых $X = x_i$, $i = 1, \ldots, n$. Из W_i случайным образом выбирается индивидуум, у которого измеряется $Y = y_i$, $i = 1, \ldots, n$. При таком подходе только Y является случайной величиной, значения X определяются условиями наблюдений.

При втором методе получения выборки, мы случайным образом отбираем n индивидуумов из W и у каждого из них измеряем как переменные X, так и Y. Здесь случайными являются обе величины X и Y. Преимущество этого метода получения выборки заключается в том, что мы можем сделать статистические выводы относительно коэффициента корреляции между X и Y, в то время как при первом методе этого сделать нельзя.

Независимо от способа получения выборки, имеются два предварительных шага для определения существования и степени линейной зависимости между X и Y. Первый шаг заключается в графическом отображении точек (x_1, y_1), (x_2, y_2), …, (x_n, y_n) на плоскость XY. Такой график называется диаграммой рассеяния. Анализируя диаграмму рассеяния, мы можем эмпирически решить, допустимо ли предположение о линейной зависимости между X и Y. Вторым шагом является вычисление выборочного коэффициента корреляции:

$$r = \frac{\sum_{i=1}^{n}(x_i - \bar{x})(y_i - \bar{y})}{\sqrt{\sum_{i=1}^{n}(x_i - \bar{x})^2 \sum_{i=1}^{n}(y_i - \bar{y})^2}}.$$

Если абсолютная величина коэффициента корреляции велика, это обоснованно указывает на сильную линейную зависимость между переменными.

1.1.1 Простая линейная регрессионная модель и оценивание по методу наименьших квадратов

Если предполагается линейная зависимость между Y и X, то теоретическая модель задается уравнениями

$$y_i = \beta_0 + \beta_1 x_i + e_i,\ i = 1, \ldots, n$$

и называется *моделью простой линейной регрессии* Y по X. Величины β_0 и β_1 являются неизвестными параметрами, а $e_1, e_2, \ldots, e_n$ ошибки, называемые отклонениями.

Оценки значений b_0 и b_1 для β_0 и β_1 по имеющейся выборке объема n можно получить минимизацией по β_0 и β_1 *суммы квадратов отклонений*:

$$S = \sum_{i=1}^{n}(y_i - \beta_0 + \beta_1 x_i)^2,$$

которая является мерой ошибки, возникающей при аппроксимации выборки прямой. Оценки b_0 и b_1 минимизируют эту ошибку.

Оценкой *уравнения регрессии* (или *прямой наименьших квадратов*) будет:

$$y' = b_0 - b_1 x.$$

Такой метод нахождения оценок называется *методом наименьших квадратов*, а регрессия называется *среднеквадратической*.

Оценки для метода наименьших квадратов задаются формулами:

$$b_0 = \bar{y} - b_1 \bar{x},$$

$$b_1 = \frac{\sum_{i=1}^{n}(x_i - \bar{x})y_i}{\sum_{i=1}^{n}(x_i - \bar{x})^2} = \frac{\sum_{i=1}^{n}(x_i - \bar{x})(y_i - \bar{y})}{\sum_{i=1}^{n}(x_i - \bar{x})^2}.$$

Оценки по методу наименьших квадратов неустойчивы к нарушениям предположения о нормальности распределения случайных ошибок (отклонений). Это связано с тем, что квадратичная мера ошибки придает слишком большой вес далеким отклонениям от регрессионной поверхности.

Возможны и другие критерии поиска подходящей аппроксимации для неизвестной истинной функции регрессии, например, минимизация *суммы модулей отклонений* (*среднеабсолютная или медианная регрессия*):

$$S = \sum_{i=1}^{n} \left| y_i - \beta_0 + \beta_1 x_i \right|,$$

или минимизация *максимального отклонения* (*минимаксная регрессия*):

$$S = \max_i \left| y_i - \beta_0 + \beta_1 x_i \right|.$$

Можно интерпретировать предсказанное значение y' двумя способами. При первом способе исследователь заинтересован в оценивании значения Y для индивидуума, у которого X принимает значение x. В этой ситуации y' есть наилучшая оценка единственного значения Y, соответствующего $X = x$. При втором подходе исследователь делает выводы о среднем значении Y для подпопуляции, соответствующей значению $X = x$. Тогда та же самая оценка y' будет наилучшей оценкой среднего значения Y при $X = x$. Различие между этими двумя способами интерпретации важно, когда строятся доверительные интервалы.

1.1.2 Доверительные интервалы и проверка гипотез

Если e_1, e_2, …, e_n — независимые случайные ошибки, имеющие нормальный закон распределения $N(0,\sigma^2)$, то можно воспользоваться следующим методом проверки гипотез и построения доверительных интервалов.

Кроме оценок параметров регрессии, нам понадобятся оценки дисперсий.

Средний квадрат отклонения (*остатка*, *ошибки*) вычисляется по формуле:

$$MS_R = SS_R / \nu_R,$$

где *остаточная сумма квадратов* (*сумма квадратов ошибок*) SS_R имеет вид:

$$SS_R = \sum_{i=1}^{n} (y_i - y_i')^2,$$

а ν_R — *число степеней свободы остатков* (или *ошибок*):

$$\nu_R = n - p - 1,$$

где p — количество независимых переменных регрессии (в случае простой линейной регрессии $p = 1$).

Обусловленный регрессией средний квадрат определяется по формуле

$$MS_D = SS_D / \nu_D,$$

где *обусловленная регрессией сумма квадратов* SS_D имеет вид

$$SS_D = \sum_{i=1}^{n} (y_i' - \bar{y})^2,$$

а ν_D — *число степеней свободы* ($\nu_D = p$).

Полная сумма квадратов определяется по формуле

$$SS_T = \sum_{i=1}^{n} (y_i - \bar{y})^2 = SS_R + SS_D.$$

Число степеней свободы для полной суммы квадратов равно

$$\nu_T = \nu_R + \nu_D = n - 1.$$

Полная сумма квадратов SS_T, деленная на число степеней свободы ν_T, равна оценке дисперсии Y.

F-отношение равно

$$F = MS_D / MS_R.$$

Наконец отношение

$$R^2 = SS_D / SS_T$$

иногда называемое коэффициентом детерминации есть доля дисперсии Y, «объясненная» регрессией Y по X_1, ..., X_p (эта величина равна квадрату множественного коэффициента корреляции). Итак, R^2 является мерой качества подгонки, т. е. чем больше R^2, тем лучше модель аппроксимирует Y.

Для проверки гипотезы о том, что простая линейная регрессия Y по X отсутствует, т. е. гипотезы $H_0 : \beta_1 = 0$ против альтернативы $H_1 : \beta_1 \neq 0$, мы используем F-отношение. Если верна гипотеза H_0, то F имеет F-распределение с ν_D и ν_R степенями свободы. P-значение есть площадь области под кривой плотности распределения $F(\nu_D, \nu_R)$ справа от F. Мы отвергаем H_0, если P меньше, чем уровень значимости α (риском ошибиться). Если H_0 принимается, то наилучшей оценкой Y при любом $X = x$ будет среднее значение $\bar{y}$.

Также можно проверить дополнительные гипотезы и построить доверительные интервалы. Для проверки $H_0 : \beta_1 = \beta_1^{(0)}$, где $\beta_1^{(0)}$ — константа, для простой линейной регрессии используем статистику:

$$t_0 = \frac{b_1 - \beta_1^{(0)}}{\sqrt{V(b_1)}},$$

где

$$V(b_1) = \frac{MS_R}{\sum_{i=1}^{n} (x_i - \bar{x})^2}.$$

Величина $\sqrt{V(b_1)}$ часто называется *стандартной ошибкой коэффициента регрессии*. Если гипотеза H_0 верна, то t_0 имеет t-распределение Стьюдента с $\nu_R = n - 2$ степенями свободы. Соответственно 100(1–α) *%-ный доверительный интервал* для β_1 есть

$$b_1 \pm \sqrt{V(b_1)}\, t_{1-(\alpha/2)}(n-2)$$

Для проверки $H_0 : \beta_0 = \beta_0^{(0)}$, где $\beta_0^{(0)}$ — константа, для простой линейной регрессии используем статистику:

$$t_0 = \frac{b_0 - \beta_0^{(0)}}{\sqrt{V(b_0)}},$$

где

$$V(b_0) = \frac{MS_R \sum_{i=1}^{n} x_i^{\,2}}{n\sum_{i=1}^{n}(x_i - \bar{x})^2}.$$

Величина $\sqrt{V(b_0)}$ иногда называется *стандартной ошибкой свободного члена*. Если гипотеза H_0 верна, то t_0 имеет t-распределение Стьюдента с $\nu_R = n - 2$ степенями свободы. Соответственно 100(1–α) *%-ный доверительный интервал* для β_0 есть:

$$b_0 \pm \sqrt{V(b_0)}\, t_{1-(\alpha/2)}(n-2).$$

Приведем теперь два доверительных интервала, основанных на оценке y'. Если $y' = b_0 - b_1 x$. интерпретируется как оценка единственного значения Y при $X = x$, то 100(1–α) *%-й доверительный интервал* для Y определяется выражением:

$$y' \pm \sqrt{MS_R}\sqrt{1 + \frac{1}{n} + \frac{(x-\bar{x})^2}{\sum_{i=1}^{n}(x_i - \bar{x})^2}}\, t_{1-(\alpha/2)}(n-2).$$

Если, с другой стороны, y' интерпретируется как оценка среднего значения Y при заданном значении $X = x$, то 100(1–α) *%-ный доверительный интервал* есть

$$y' \pm \sqrt{MS_R}\sqrt{\frac{1}{n} + \frac{(x-\bar{x})^2}{\sum_{i=1}^{n}(x_i - \bar{x})^2}}\, t_{1-(\alpha/2)}(n-2).$$

Выбор доверительного интервала зависит от того, как используется оценка y' исследователем. Заметим, что когда x удаляется от $\bar{x}$, доверительный интервал увеличивается, т.е. оценка становится менее точной.

1.1.3 Проверка адекватности линейной модели

Теперь мы обсудим, каким образом проверить адекватность модели линейной регрессии. Под адекватностью модели линейной регрессии подразумевается, что никакая другая линейная модель не даст значимого улучшения в предсказании *Y*. Пусть, например, исследователь пожелал проверить, значимо ли улучшается предсказание *Y* с помощью *модели полиномиальной регрессии* $y_i = \beta_0 + \beta_1 x_i + \beta_2 x_i^2 + \ldots + \beta_m x_i^m + e_i$ для некоторого $m \geq 2$. Нулевой гипотезой в этом случае будет $H_0 : \beta_2 = \ldots = \beta_m = 0$.

Если для некоторых значений из *X* имеется более чем по одному значению из *Y*, то можно проверить гипотезу, что никакая альтернативная модель не дает значимого улучшения предсказания *Y* по сравнению с моделью простой линейной регрессии. Статистика критерия есть еще одно *F*-отношение, которое получается из таблицы дисперсионного анализа следующим образом.

Предположим, что имеется *k* различных значений для *X*, например x_1, …, x_k. Далее, предположим, что для каждого из этих x_i имеется n_i наблюдений $y_{i1}, y_{i2}, \ldots, y_{in_i}$ переменной *Y*, $i = 1, \ldots, k$. Пусть $n_i > 1$ для некоторого *i*, и пусть $\sum_{i=1}^{k} n_i = n$. Тогда модель простой линейной регрессии может быть записана в следующем виде:

$$y_{ij} = \beta_0 + \beta_1 x_i + e_{ij}, \; i = 1, \ldots, k, \; j = 1, \ldots, n_i.$$

Пусть e_{ij} независимые случайные величины, распределенные по закону $N(0,\sigma^2)$.

Можно получить оценки b_0 и b_1 для β_0 и β_1 обрабатывая выборку *n* двумерных наблюдений $(x_1, y_{11}), (x_1, y_{12}), \ldots, (x_1, y_{1n_1}), \ldots, (x_k, y_{k1}), (x_k, y_{k2}), \ldots, (x_k, y_{kn_k})$. Эти оценки имеют вид:

$$b_0 = \bar{y}.. - b_1 \bar{x},$$

$$b_1 = \frac{\sum_{i=1}^{n} n_i (x_i - \bar{x}) \bar{y}_i .}{\sum_{i=1}^{k} n_i (x_i - \bar{x})^2},$$

где

$$\bar{y}_i . = \frac{1}{n_i} \sum_{j=1}^{n_i} y_{ij}, \quad \bar{y}.. = \frac{1}{n} \sum_{i=1}^{k} \sum_{j=1}^{n_i} y_{ij} \text{ и } \bar{x} = \frac{1}{n} \sum_{i=1}^{k} n_i x_i .$$

Прямая наименьших квадратов есть $y' = b_0 - b_1 x$, так что $y_i' = b_0 - b_1 x_i$ есть оценка *Y* при $X = x_i$.

Суммами квадратов являются

$$SS_D = \sum_{i=1}^{k}\sum_{j=1}^{n_i}(y_i' - \bar{y}..)^2 \text{ и } SS_R = \sum_{i=1}^{k}\sum_{j=1}^{n_i}(y_{ij} - y_i')^2$$

с $\nu_D = 1, \nu_R = n - 2$ степенями свободы соответственно.

Для проверки гипотезы об адекватности линейной модели остаточная сумма квадратов SS_R и число степеней свободы ν_R делятся между двумя источниками дисперсии относительно регрессии и внутри групп. Соответствующие суммы квадратов SS_A и SS_W и степени свободы ν_A и ν_W будут равны

$$SS_A = \sum_{i=1}^{k}\sum_{j=1}^{n_i}(\bar{y}. - y_i')^2,\ \nu_A = k - 2,$$

$$SS_W = \sum_{i=1}^{k}\sum_{j=1}^{n_i}(y_{ij} - \bar{y}_i.)^2,\ \nu_W = n - k,$$

$$SS_R = SS_A + SS_W,\ \nu_R = \nu_A + \nu_W.$$

Статистика критерия для проверки гипотезы H_0 : «простая линейная модель адекватна», против H_1 : «простая линейная модель неадекватна», есть

$$F = MS_A / MS_W,$$

где MS_A и MS_W — соответственно средние квадраты разброса относительно регрессии и внутри групп:

$$MS_A = SS_A / \nu_A,$$
$$MS_W = SS_W / \nu_W.$$

В случае истинности H_0 величина F_0 имеет распределение $\nu_A = k - 2$ и $\nu_W = n - k$ степенями свободы. P-значение есть площадь области под кривой плотности распределения $F(\nu_A, \nu_W)$ справа от F_0.

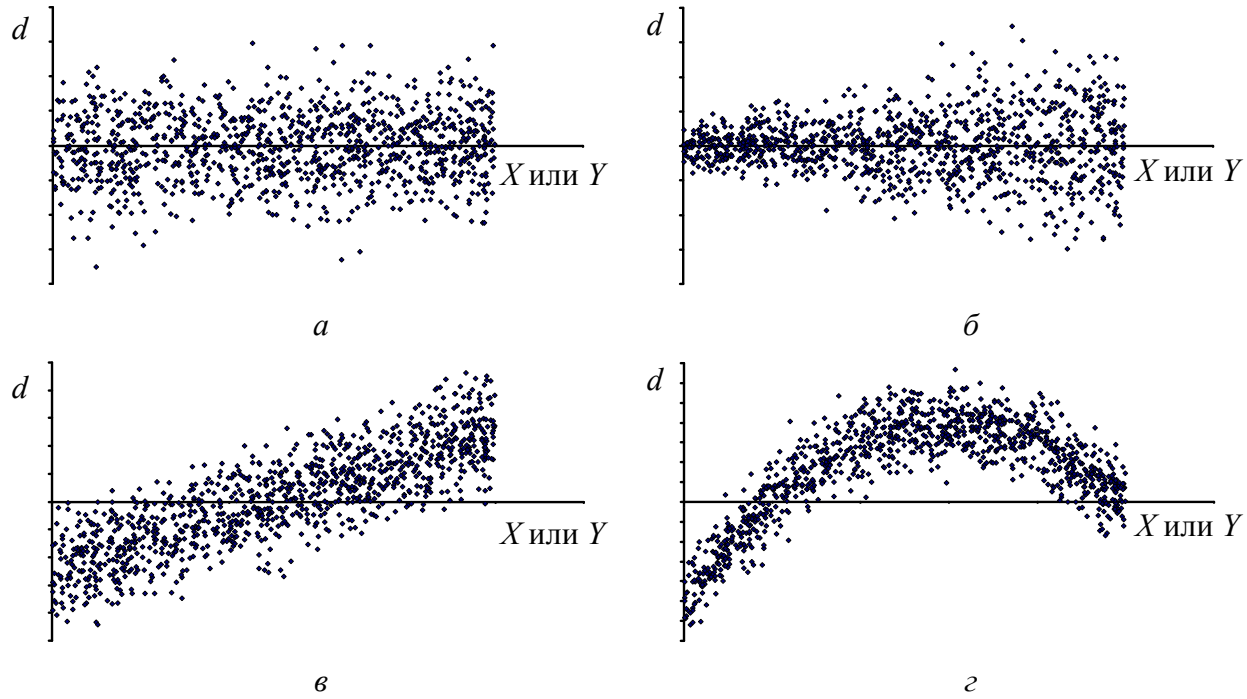
Если H_0 принимается, то после этого с помощью F-отношения, заданного ранее, может быть проверена гипотеза $H_0 : \beta_1 = 0$.

1.1.4 Анализ остатков

В предыдущем рассмотрении простой линейной регрессии были сделаны три предположения. Они касались формы модели, распределения и случайности величины ошибки e.

Все три предположения могут быть проверены при рассмотрении графиков остатков $d_i = y_i - y_i'$, $i = 1, \ldots, n$. Для проверки адекватности модели можно использовать график d_i в зависимости от x_i или y_i' $i = 1, \ldots, n$. Если остатки попадают в горизонтальную полосу с центром на оси абсцисс, модель можно рассматривать как адекватную (рисунок 1, *а*). Если полоса расширяется (сужается), когда x или y_i' возрастает (рисунок 1, *б*), это указывает на *гетероскедастичность* (т. е. на отсутствие постоянства дисперсии σ^2). В частности, σ может быть функцией $\beta_0 + \beta_1 x$, что делает необходимым преобразование переменной Y. График, показывающий линейный тренд (рисунок 1 *в*), дает основание для введения в модель дополнительной независимой перемен-

ной. График вида, представленного на рисунке 1 *г*, указывает, что в модель должен быть добавлен линейный и квадратичный член.

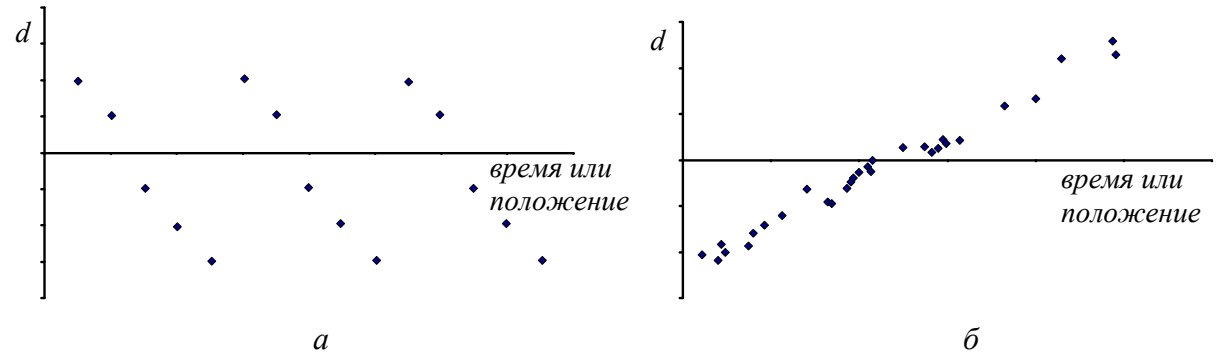


а — адекватная модель; *б* — гетероскедастичность;
в — линейная независимая переменная;
г — линейная и квадратичная независимая переменная

Рисунок 1 — Примеры графиков остатков

Для проверки нормальности e_i, $i = 1, \ldots, n$, подходит гистограмма d_i. Нормальность может быть также проверена с помощью критериев согласия.

Если данные упорядочены некоторым образом (например, последовательность точек по времени или по расположению), то график остатков d_i в том же самом порядке, в котором собирались данные, позволяет проверить случайность. Гипотезу о случайности можно отвергнуть, если выявлен тренд, причем тренд может иметь как сезонный, так и линейный характер (рисунок 2 *а* и *б*)



а — сезонный тренд, *б* — линейный тренд

Рисунок 2 — Примеры отсутствия случайности

1.2 Множественная линейная регрессия

Рассмотрим теперь проблему предсказания одной переменной Y с помощью p переменных $X_1, X_2, \ldots, X_p$, $p > 1$. Традиционно переменная Y называется зависимой переменной, в то время как переменные $X_1, \ldots, X_p$ называются независимыми переменными. Такое применение слова «независимые» не следует смешивать с понятием «статистической независимости». Фактически, в некоторых случаях независимые переменные $X_1, \ldots, X_p$ суть случайные величины, которые не обязательно являются статистически независимыми.

Величину Y можно аппроксимировать посредством функции регрессии $f(\cdot)$, содержащей неизвестные параметры. Уравнение модели, выражающей зависимость между зависимой и независимыми переменными, можно записать в виде

$$y = f(x_1,\ldots,x_p;\beta_1,\ldots,\beta_m) + e,$$

где $\beta_1, \ldots, \beta_m$ — неизвестные параметры и e — ошибка аппроксимации Y посредством функции регрессии. В частности, если $m = p + 1$ и $f(x_1,\ldots,x_p;\beta_0,\beta_1,\ldots,\beta_p) = \beta_0 + \beta_1 x_1 + \ldots + \beta_p x_p$, мы имеем модель множественной линейной регрессии

$$y = \beta_0 + \beta_1 x_1 + \ldots + \beta_p x_p + e.$$

В этом уравнении некоторые независимые переменные могут быть функциями других переменных или друг друга. Например, $y = \beta_0 + \beta_1 \sin z_1 + \beta_2 \cos z_2 + e$ есть модель множественной линейной регрессии с $x_1 = \sin z_1$ и $x_2 = \cos z_2$. В частности, если $x_i = z^i$, $i = 1, \ldots, p$, получается модель полиномиальной регрессии

$$y = \beta_0 + \beta_1 x + \beta_1 x^2 + \ldots + \beta_p x^p + e.$$

Нужно помнить, что слово «линейная» подразумевает линейность относительно параметров, но не по отношению к независимым переменным. Так, $y = \beta_0 + \sin(\beta_1 x_1) + \beta_2 x_2$ не является линейной функцией параметров.

Хотя для описания многих реальных ситуаций более подходящими являются нелинейные модели, линейная модель может быть полезна, по крайней мере, как первое приближение к нелинейной модели.

1.2.1 Оценивание параметров

Параметры модели оцениваются по выборке объема n, полученной из популяции W. Так же как и ранее, эту выборку можно получить одним из двух способов. При первом способе фиксируются некоторые значения $X_1, \ldots, X_m$, а затем в подпопуляции, определенной этими ограничениями, наблюдаются одно или несколько значений переменной Y. Затем фиксируются новые значения $X_1, \ldots, X_p$ и наблюдаются одно или несколько значений Y в этой подпопуляции, и так продолжается до тех пор, пока не будет получено n наблюдений. При таком способе формирования выборки случайной является лишь переменная Y. Второй способ получения выборки заключается в случайном отборе n индивидуумов из популяции W и одновременном наблюде-

нии у них всех $p + 1$ переменных Y, X_1, …, X_p, причем все эти переменные случайны. Хотя процедура оценивания параметров одинакова для всех способов формирования выборки, одно из основных предположений теории оценивания методом наименьших квадратов состоит в том, что выборка образована первым способом. С другой стороны, теория множественного и частного коэффициентов корреляции основывается на том, что выборка образована по второму способу из многомерной нормальной популяции.

Здесь предполагается, что x_{1i}, …, x_{pi}, $i = 1, \ldots, n$, суть фиксированные значения независимых переменных X_1, …, X_p (здесь $X_1 = x_{1i}$, …, $X_p = x_{pi}$, а y_i — наблюдаемое значение переменной Y). Итак, выборка состоит из n наблюдений (y_1; x_{11}, …, x_{p1}), …, (y_n; x_{1n}, …, x_{pn}). Для модели множественной линейной регрессии имеем

$$y_i = \beta_0 + \beta_1 x_{1i} + \ldots + \beta_p x_{pi} + e_i,$$

где β_0, β_1, …, β_p — неизвестные параметры, а e_1, …, e_n — независимые случайные ошибки, распределенные по закону $N(0, \sigma^2)$. Оценки параметров b_0, b_1, …, b_p, которые минимизируют *сумму квадратов отклонений*

$$S = \sum_{i=1}^{n} (y_i - \beta_0 - \beta_1 x_{1i} - \ldots - \beta_p x_{pi})^2,$$

обычно называются (*частными*) *коэффициентами регрессии*. Иногда оценка b_0 называется *свободным членом*, *константой* или *смещением* по y. Оценка уравнения множественной линейной регрессии (или *плоскость наименьших квадратов*) может быть записана в виде

$$y' = b_0 + b_1 x_1 + \ldots + b_p x_p.$$

Заметим, что сумма квадратов отклонений S является мерой ошибки, связанной с «подгонкой» выборочных данных посредством модели линейной регрессии; МНК-оценки минимизируют эту ошибку. Далее, b_i суть несмещенные оценки для β_i, $i = 0, 1, \ldots, p$, и выражаются линейными функциями наблюдений y_1, …, y_n. Наконец, из теоремы Гаусса—Маркова следует, что предсказанное значение y имеет минимальную дисперсию для данных x_1, …, x_p среди всех линейных по X_1, …, X_p предикторов Y.

1.2.2 Проверка гипотез

Для проверки гипотезы о том, что множественная линейная регрессия отсутствует, т. е. гипотезы $H_0 : \beta_1 = \ldots = \beta_p = 0$, которую можно рассматривать как гипотезу о том, что «независимые переменные X_1, …, X_p не улучшают предсказание Y относительно $y' = \bar{y}$», против альтернативой гипотезы, что не все коэффициенты равны нулю, мы используем F-отношение:

$$F = MS_D / MS_R.$$

Статистика F для H_0 имеет F-распределение с $\nu_D = p$ и $\nu_R = n - p - 1$ степенями свободы. Соответствующее P-значение есть площадь области F под кривой плотности распределения $F(\nu_D, \nu_R)$ справа от точки, соответствующей вычисленному значению F.

Труднее проверить промежуточную гипотезу о равенстве нулю некоторого подмножества из m коэффициентов. Без потери общности предположим, что подмножество состоит из первых m коэффициентов $\beta_1, \ldots, \beta_m$. Тогда проверка гипотезы $H_0 : \beta_1 = \ldots = \beta_m = 0$ эквивалентна проверке гипотезы о том, что «m переменных $X_1, \ldots, X_m$ не улучшают предсказание Y относительно предсказания, получаемого с помощью регрессии Y по $X_1, \ldots, X_p$». Для проверки H_0 сначала вычислим регрессию Y по переменным $X_{m+1}, \ldots, X_p$ и найдем остаточную сумму квадратов SS'_R. Затем вычислим регрессию Y по всему набору переменных $X_1, \ldots, X_m, \ldots, X_p$. Остаточную сумму квадратов и средний квадрат для этого случая обозначим через SS_R и MS_R соответственно. Тогда статистика критерия для H_0 имеет вид

$$F = \frac{(SS'_R - SS_R)/m}{MS_R}.$$

Для гипотезы H_0 она имеет F-распределение с m и $\nu_R = n - p - 1$ степенями свободы. P-значение есть площадь области, расположенной под кривой плотности распределения $F(m, \nu_R)$ справа от точки F, равной вычисленному значению F.

Также можно проверить гипотезу $H_0 : \beta_k = \beta_k^{(0)}$, где $\beta_k^{(0)}$ — заданная константа, а также построить доверительные интервалы.

1.2.3 Дополнение к анализу остатков

Ранее рассматривалось использование графиков остатков $d_i = y_i - y'_i$ в зависимости от x_i или y'_i ($i = 1, \ldots, n$) для проверки предположений модели простой линейной регрессии. Аналогичные графики могут быть построены и в случае модели множественной линейной регрессии. Однако здесь можно получить значительно больше графиков, поскольку остатки можно сопоставлять с каждой из p независимых переменных.

График d_i в сопоставлении с x_{ji} ($i = 1, \ldots, n$, $j = 1, \ldots, p$) содержит информацию о:

а) наличии аномальных наблюдений или случаев отклонений по j-й независимой переменной;

б) возможном отсутствии линейности по X_j, что может служить указанием для дальнейшего преобразования.

График d_i относительно y_i ($i = 1, \ldots, n$) доставляет информацию о выполнении предположений случайности и независимости ошибок e_i, а также и предположения о гомоскедастичности e_i.

1.3 Нелинейная регрессия

Ранее рассматривались модели регрессии, линейные по параметрам, вида

$$y = \beta_0 + \beta_1 x_1 + \ldots + \beta_p x_p + e, \; i = 1, \ldots, n.$$

Во многих случаях линейная модель может служить, по меньшей мере, в качестве первого приближения к истинной модели. Кроме того, в некоторых случаях использование подходящих преобразований переменных может привести к линейной по парамётрам модели. Однако имеется большое число ситуаций, для которых линейная модель непригодна, например, когда зависимость выражается суммой экспоненциальных и/или тригонометрических функций. В этом случае линейная модель не будет уже удовлетворительной аппроксимацией, а простое преобразование переменных, приводящее к ней, отсутствует.

Любая модель, вид которой не совпадает с приведенным уравнением, называется моделью нелинейной регрессии и может быть представлена в виде

$$y_i = f(x_{1i},\ldots,x_{pi};\theta_1,\ldots,\theta_m) + e_i,\ i = 1,\ \ldots,\ n,$$

где $f(\cdot)$ — нелинейная функция параметров $\theta_1,\ \ldots,\ \theta_m$, а e_i — некоррелированные ошибки. Приведем два примера нелинейной функции:

$$f(x_i;\theta_1,\theta_2,\theta_3) = \theta_1 + \theta_2 e^{\theta_3 x_i},$$

$$f(x_{1i},x_{2i};\theta_1,\theta_2,\theta_3) = \theta_1 + \theta_2 \sin(x_{1i} + \theta_3 \cos x_{2i}),\ i = 1,\ \ldots,\ n.$$

Если истинная модель линейна, то МНК-оценки параметров будут оптимальными, поскольку они являются несмещенными оценками с минимальной дисперсией. Но если модель нелинейна, то методы получения наилучших оценок параметров отсутствуют.

Однако существует метод *максимального правдоподобия*, который позволяет получать оценки $\theta'_1,\theta'_2,\ldots,\theta'_m$, обладающие такими ценными свойствами, как *состоятельность* и *асимптотическая эффективность* при достаточно общих условиях. Более того, если ошибки e_i суть независимые случайные величины с распределением $N(0,\sigma^2)$, оценки максимального правдоподобия совпадают с МНК-оценками. МНК-оценки суть значения $\theta_1,\ \theta_2,\ \ldots,\ \theta_m$, которые минимизируют *сумму квадратов отклонений*:

$$S = \sum_{i=1}^{n} (y_i - f(x_{1i},\ldots,x_{pi};\theta_1,\ldots,\theta_m))^2 .$$

Для линейной модели МНК-оценки получаются из решения системы линейных уравнений. К сожалению, в случае нелинейной модели приходится решать систему нелинейных уравнений и соответствующее МНК-решения нельзя уже представить в явном виде. По этой причине приходится использовать различные итерационные методы для численного определения МНК-оценок.

2 ПОРЯДОК ВЫПОЛНЕНИЯ РАБОТЫ

Задача 1. Калибруется датчик для измерения давления в пневматической тормозной системе. Исследователь выполнил n = 20 измерений (выборок) с давления с помощью эталонного прибора и калибруемого датчика. Пусть X обозначает давление, измеренное с помощью эталонного прибора (кПа), а Y — давление (кПа), определенное с помощью датчика. Полученные данные приведены в таблице 1 (файл «Калибровка.xls»).

Таблица 1 – Данные калибровки прибора для измерения концентрации молочной кислоты в крови

X	***Y***	***X***	***Y***	***X***	***Y***	***X***	***Y***
50	55	150	70	250	410	500	660
50	35	150	245	250	310	750	935
50	90	150	220	500	600	750	985
50	20	150	225	500	655	750	870
150	150	250	365	500	630	750	855

Заметим, что эти данные относятся к первому способу формирования выборки, так что X фиксировано на уровне одного из пяти значений: X = 50, X = 150, X = 250, X = 500 или X = 750.

Рассчитаем выборочный коэффициент корреляции. Для этого:

а) откройте файл «Калибровка.xls» и сохраните его как файл «Калибровка 1.xls».

б) введите в ячейку A22 формулу для расчета $\bar{x}$ «=СРЗНАЧ(A2:A21)» .

в) введите в ячейку B22 формулу для расчета $\bar{y}$ «=СРЗНАЧ(B2:B21)» .

г) рассчитайте в столбце C значения $(x_i - \bar{x})$ (введите в ячейку C2 формулу «=A2-A$22» и скопируйте ее в ячейки C3:C21).

д) рассчитайте в столбце D значения $(y_i - \bar{y})$ (выделите ячейки C2:C21 и скопируйте их в ячейки D2:D21).

е) рассчитайте в столбце E значения $(x_i - \bar{x})^2$ (введите в ячейку E2 формулу «=C2^2» и скопируйте ее в ячейки E3:E21).

ж) рассчитайте в столбце F значения $(y_i - \bar{y})^2$ (выделите ячейки E2:E21 и скопируйте их в ячейки F2:F21).

з) рассчитайте $\sum_{i=1}^{n}(x_i - \bar{x})^2$. Для этого посчитайте в ячейке E22 сумму по ячейкам E2:E21).

и) рассчитайте $\sum_{i=1}^{n}(y_i - \bar{y})^2$. Для этого посчитайте в ячейке F22 сумму по ячейкам F2:F21).

к) рассчитайте в столбце G значения $(x_i - \bar{x})(y_i - \bar{y})$ (введите в ячейку G2 формулу «=C2*D2» и скопируйте ее в ячейки G3:G21).

л) рассчитайте $\sum_{i=1}^{n}(x_i - \bar{x})(y_i - \bar{y})$. Для этого посчитайте в ячейке G22 сумму по ячейкам G2:G21).

м) рассчитайте выборочный коэффициент корреляции. Для этого введите в ячейку H22 формулу «=G22/КОРЕНЬ(E22*F22)».

Выборочный коэффициент корреляции r = 0,987 близок к единице, что указывает на очень сильную линейную зависимость между X и Y.

Построим диаграмму рассеяния. Для этого:

а) выделите ячейки A2:B21.
б) выберите команду меню «Вставка ⇒ Диаграмма…».
в) выберите тип диаграммы «Точечная» и нажмите «Далее».
г) нажмите кнопку «Готово».

На диаграмме рассеяния также прослеживается линейный характер зависимости между X и Y.

В Excel есть две функции для расчета оценок коэффициентов простой линейной модели на основе метода наименьших квадратов (таблица 2).

Таблица 2 – Функции Excel для оценки регрессии на основе метода наименьших квадратов

Функция	Описание
ОТРЕЗОК (y, x)	Возвращает точку пересечения линии регрессии с осью Y по известным значениям x и y
НАКЛОН (y, x)	Возвращает наклон линии регрессии по известным значениям x и y

Найдем оценку коэффициентов регрессионного уравнения. Для этого:

а) введите в ячейку C24 формулу для расчета коэффициента регрессии b_0 «=ОТРЕЗОК(B2:B21;A2:A21)».

б) введите в ячейку C25 формулу для расчета коэффициента регрессии b_1 «=НАКЛОН(B2:B21;A2:A21)».

В результате получим уравнение простой линейной регрессии y' = 7,974 + 1,228 x. Для практических целей желательно предсказать истинную концентрацию X по наблюдаемой концентрации Y. Для этого нужно обратить оценку регрессионного уравнения, что дает для оценки X по Y уравнение x = (y – 7,974)/1,228.

Построим линию регрессии графически на ранее построенной диаграмме рассеяния. Для этого:

а) щелкните на диаграмме рассеяния правой кнопкой мыши на любую точку диаграммы.

б) выберите команду в появившемся контекстном меню команду «Добавить линию тренда …».

в) выберите вкладку «Тип» и выберите тип линии тренда «Линейная».

г) выберите вкладку «Параметры» и установите флажок «показывать уравнение на диаграмме».

д) щелкните на кнопке «ОК».

Построим таблицу дисперсионного анализа. Для этого:

а) рассчитайте в столбце I значения $y'_i = b_0 + b_1 x_i$ (введите в ячейку I2 формулу «=C$24+C$25*A2» и скопируйте ее в ячейки I3:I21).

б) рассчитайте в столбце J значения $(y'_i - \bar{y})^2$ (введите в ячейку J2 формулу «=(I2-B$22)^2» и скопируйте ее в ячейки J3:J21).

в) рассчитайте в столбце K значения $(y_i - y'_i)^2$ (введите в ячейку K2 формулу «=(B2-I2)^2» и скопируйте ее в ячейки K3:K21).

г) рассчитайте в ячейке G24 обусловленную регрессией сумму квадратов SS_D (введите в ячейку формулу «=СУММ(J2:J21)»).

д) рассчитайте в ячейке G25 остаточную сумму квадратов SS_R (введите в ячейку формулу «=СУММ(K2:K21)»).

е) рассчитайте в ячейке G26 полную сумму квадратов SS_T (введите в ячейку формулу «=G24+G25», если все сделано правильно полученное значение должно совпадать со значением в ячейке F22).

ж) введите в ячейку H24 число степеней свободы обусловленной регрессией суммы квадратов ν_D = 1 (количество коэффициентов регрессии минус один).

з) рассчитайте в ячейке H26 число степеней свободы для полной суммы квадратов $\nu_T = n - 1$ (введите формулу «=СЧЁТ(B2:B21)-1»).

и) рассчитайте в ячейке H25 число степеней свободы остатков $\nu_R = \nu_T - \nu_D$ (введите формулу «=H26-H24»).

к) рассчитайте в ячейке I24 обусловленный регрессией средний квадрат $MS_D = SS_D / \nu_D$ (введите формулу «=G24/H24»).

л) рассчитайте в ячейке I25 средний квадрат отклонения $MS_R = SS_R / \nu_R$ (введите формулу «=G25/H25»).

м) рассчитайте в ячейке J24 F-отношение $F = MS_D / MS_R$ (введите формулу «=I24/I25»).

Проверим гипотезу о том, что простая линейная регрессия Y по X не улучшает оценку Y по сравнению со средним значением Y, т. е. гипотезу H_0 : $\beta_1 = 0$ против альтернативы H_1 : $\beta_1 \neq 0$. Для этого рассчитайте в ячейке K24 P-

значение равное площади области под кривой плотности распределения $F(\nu_D, \nu_R)$ справа от F (введите формулу «=FРАСП(J24;H24;H25)»).

Поскольку P-значение очень мало гипотеза не принимается.

Проверим гипотезу, что прямая регрессии проходит через начало координат, т.е. гипотезу $H_0 : \beta_0 = 0$ против альтернативы $H_1 : \beta_0 \neq 0$. Для этого построим 95%-й доверительный интервал для β_0:

а) рассчитайте в столбце L значения x_i^2 (введите в ячейку L2 формулу «=A2^2» и скопируйте ее в ячейки L3:L21).

б) рассчитайте в ячейке L22 $\sum_{i=1}^{n} x_i^2$ (введите в ячейку формулу «=СУММ(L2:L21)»).

в) рассчитайте в ячейке A28 стандартную ошибку свободного члена $\sqrt{V(b_0)}$ (введите формулу «=КОРЕНЬ(I25*L22/СЧЁТ(A2:A21)/E22)»).

г) рассчитайте в ячейке B28 статистику $t_0 = (b_0 - \beta_0^{(0)}) / \sqrt{V(b_0)}$ (введите формулу «=(C24-0)/A28»).

д) рассчитайте в ячейке C28 уровень значимости, с которым мы можем отвергнуть гипотезу (введите формулу «=1-СТЬЮДРАСП(B28; СЧЁТ(A2:A21)-2;2)»).

е) рассчитайте в ячейке D28 левую границу доверительного интервала для b_0 (введите формулу «=C24-A28*СТЬЮДРАСПОБР(0,05; СЧЁТ(A2:A21)-2)»).

ж) рассчитайте в ячейке E28 правую границу доверительного интервала для b_0 (введите формулу «=C24+A28*СТЬЮДРАСПОБР(0,05; СЧЁТ(A2:A21)-2)»).

Так как этот интервал включает нуль, гипотеза H_0 не отвергается.

Найдем 95%-й интервала для среднего значения Y при $X = 385$. Для этого:

а) найдите оценку среднего значения Y при $X = 385$, равную $y' = b_0 + b_1 \cdot 385$ (введите в ячейку A31 формулу «=C24+C25*385»).

б) рассчитайте в ячейке B31 полуширину доверительного интервала $\sqrt{MS_R}\sqrt{1/n + (x - \bar{x})^2 / \sum_{i=1}^{n}(x_i - \bar{x})^2}\, t_{1-(\alpha/2)}(n-2)$ (введите формулу «=КОРЕНЬ(I25*(1/СЧЁТ(A2:A21)+(385-22)^2/E22))*СТЬЮДРАСПОБР(0,05; 18))»).

в) рассчитайте в ячейке C31 левую границу доверительного интервала (введите формулу «=A31-B31»).

г) рассчитайте в ячейке D31 правую границу доверительного интервала (введите формулу «=A31+B31»).

Полученный доверительный интервал включает истинное среднее значение Y при $X = 385$ с доверительным уровнем 95%.

Поскольку для значений из X имеется более чем по одному значению из Y, то проверим гипотезу, что никакая альтернативная модель не дает значимого улучшения предсказания Y по сравнению с моделью простой линейной регрессии. Для этого:

а) введите в ячейки A33:A37 различные значения X равные 50, 150, 250, 500, 750.

б) введите в ячейку B33 формулу «="="&ТЕКСТ(A33;"0")» (задает условие в виде текстовой строки).

в) введите в ячейку C33 формулу «=СЧЁТЕСЛИ(A$2:A$21;B33)» (подсчет количества значений Y для заданного X).

г) введите в ячейку D33 формулу «=СУММЕСЛИ(A$2:A$21;B33; B$2:B$21)/C33» (подсчет среднего значения Y для заданного X).

д) введите в ячейку E33 формулу «=C$24+C$25*A33» (подсчет $y'_i = b_0 + b_1 x_i$).

е) введите в ячейку F33 формулу «=(E33-D33)^2*C33» (подсчет $\sum_{j=1}^{n_i} (\bar{y}_\cdot - y'_i)^2$).

ж) скопируйте ячейку B33 в ячейки B34:B37, ячейку C33 в ячейки C34:C37, ячейку D33 в ячейки D34:D37, ячейку E33 в ячейки E34:E37, ячейку F33 в ячейки F34:F37.

з) рассчитайте в ячейке G33 сумму квадратов относительно отклонения от регрессии $SS_A = \sum_{i=1}^{k} \sum_{j=1}^{n_i} (\bar{y}_\cdot - y'_i)^2$ (введите формулу «=СУММ(F33:F37)»).

и) рассчитайте в ячейке G34 сумму квадратов относительно внутригруппового разброса $SS_W = SS_R - SS_A$ (введите формулу «=G25-G33»).

к) введите в ячейку H33 количество степеней свободы суммы квадратов относительно отклонения от регрессии ν_A равное количество различных значений X минус два ($k - 2 = 5 - 2 = 3$).

л) рассчитайте в ячейке H34 количество степеней свободы суммы квадратов относительно внутригруппового разброса $\nu_W = n - k$ (введите формулу «=H25-H33»).

м) рассчитайте в ячейке I33 средний квадрат $MS_A = SS_A / \nu_A$ (введите формулу «=G33/H33»).

н) рассчитайте в ячейке I34 средний квадрат $MS_W = SS_W / \nu_W$ (введите формулу «=G34/H34»).

о) рассчитайте в ячейке J33 F-отношение $F = MS_A / MS_W$ (введите формулу «=I33/I34»).

п) рассчитайте в ячейке K33 P-значение равное площади области под кривой плотности распределения $F(\nu_A,\ \nu_W)$ справа от F — уровень значимости, с которым гипотеза не отклоняется (введите формулу «=FРАСП(J33;H33;H34)»).

Уровень значимости достаточно высок, отклонить гипотезу об адекватности простой линейной модели нельзя.

Выполним анализ остатков. Для этого:

а) рассчитайте в столбце M значения $(y_i - y_i')$ (введите в ячейку M2 формулу «=(B2-I2)» и скопируйте ее в ячейки M3:M21).

б) выделите ячейки M2:M21.

в) выберите команду меню «Вставка ⇒ Диаграмма…».

г) выберите тип диаграммы «Точечная» и нажмите «Далее».

д) выберите вкладку «Ряд».

е) в поле «Значения X» выберите диапазон «=Лист1!$A$2:$A$21».

ж) нажмите кнопку «Готово».

Диаграмма остатки имеет вид скорее представленный на рисунке 1 *г*, чем на рисунке 1 *а*. Исходя из этого, стоит добавить к уравнению регрессии квадратичный член.

Задание. С помощью построения линии тренда получите полиномиальную модель степени два. Проверьте адекватность новой линейной модели и посмотрите, как изменится уровень значимости.

Задача 2. В процессе исследования свойств гидравлического масла с помощью вискозиметра была измерена кинематическая вязкость, мм^2/с, при разных температурах, ºС (файл «Вязкость.xls»).

Построим регрессионную линейную модель зависимости кинематической вязкости от температуры.

Для начала построим диаграмму рассеяния. Для этого в первую очередь необходимо отобрать из выборки данные соответствующие моменту поступления больных в больницу:

а) откройте файл «Вязкость.xls» и сохраните его под именем «Вязкость 1.xls».

б) выделите ячейки A2:B113 в новой рабочей книге (в дальнейшем будем работать только с новой рабочей книгой).

в) выберите команду меню «Вставка ⇒ Диаграмма…».

г) выберите тип диаграммы «Точечная» и нажмите «Далее».

д) нажмите кнопку «Готово».

В результате получим диаграмму рассеяния Y в сопоставлении с X, где Y есть вязкость, а X — температура.

Эта диаграмма рассеяния указывает на экспоненциальную зависимость между X и Y.

Для нефтепродуктов выведена эмпирическая формула зависимости вязкости от температуры:

$$\lg\lg(v + 0{,}6) = A + B\lg T,$$

где v – кинематическая вязкость, мм^2/с; T – температура, ºК; A и B – постоянные зависящие от свойств нефтепродукта.

Таким образом, если перевести температуру в ºК (X_K) и использовать log log (Y+0,6), можно получить линейную зависимость от log X_K. Если преобразованием переменных удается перейти к линейной зависимости, то говорят, что модель *существенно линейна*.

Задание. Перевести температуру в ºК и прологарифмируйте ее. Прибавьте к вязкости 0,6 и дважды прологарифмируйте ее. Постройте диаграмму рассеяния для преобразованной переменной. С помощью линии тренда постройте простую линейную модель. Постройте диаграмму остатков.

Помимо преобразования X и Y можно к исходным данным можно применить технику нелинейной регрессии. Воспользуемся экспоненциальной моделью вида:

$$y_i = \theta_1 + \theta_2 e^{\theta_3 x_i} + e_i.$$

Найдем значения коэффициентов θ_1, θ_2, θ_3. Для этого в первую очередь найдем сумму квадратов ошибок:

а) откройте файл «Вязкость.xls» и сохраните его под именем «Вязкость 2.xls».

б) введите в ячейку C1 начальное приближение для $\theta_1 = 1$.

в) введите в ячейку D1 начальное приближение для $\theta_2 = 10$.

г) введите в ячейку E1 начальное приближение для $\theta_3 = -0{,}1$.

д) рассчитайте в столбце C значения $y_i' = \theta_1 + \theta_2 e^{\theta_3 x_i}$ (введите в ячейку G2 формулу «=C$1+D$1*EXP(E$1*A2))» и скопируйте ее в ячейки C3:C16).

е) рассчитайте в столбце D значения ошибок (остатков) $(y_i - y_i')$ (введите в ячейку D2 формулу «=C2-B2» и скопируйте ее в ячейки D3:D16).

ж) рассчитайте в столбце E значения квадрата ошибок $(y_i - y_i')^2$ (введите в ячейку E2 формулу «=D2^2» и скопируйте ее в ячейки E3:E16).

з) рассчитайте в ячейке E17 сумму квадратов ошибок (введите формулу «=СУММ(E2:E16)»).

Для поиска коэффициентов θ_1, θ_2, θ_3 необходимо минимизировать сумму квадратов ошибки по этим коэффициентам.

В Excel для решения задач оптимизации используется подключаемый модуль (надстройка) «Поиск решения».

Для загрузки подключаемого модуля «Поиск решения» выполните перечисленные ниже действия:

а) выберите команду меню «Сервис ⇒ Надстройки».

б) установите флажок для элемента «Поиск решения» и щелкните на кнопке ОК.

При необходимости укажите путь до установочного пакета Excel.

После инсталляции и загрузки подключаемый модуль «Поиск решения» будет доступен из меню Excel в виде новой команды меню «Сервис ⇒ Поиск решения…».

Найдем значения коэффициентов θ_1, θ_2, θ_3. Для этого:

в) выделите ячейку E17, содержащую сумму квадратов ошибки.

г) выберите команду меню «Сервис ⇒ Поиск решения…».

д) на экране появится диалоговое окно «Поиск решения», которое показано на рисунке 3.

е) проверьте целевую ячейку, она должна быть равна «$E$17».

ж) выберите вариант оптимизации — минимизация (выберите «Равной: минимальному значению»).

з) укажите ячейки, значения которых будут изменяться в процессе поиска решения (ячейки $C$1:$E$1, содержащие значения коэффициентов θ_1, θ_2, θ_3).

и) нажмите кнопку «Выполнить».

к) выберите «Сохранить найденное решение».

л) нажмите кнопку «ОК».

Задание. Постройте диаграмму остатков. Сделайте выводы.

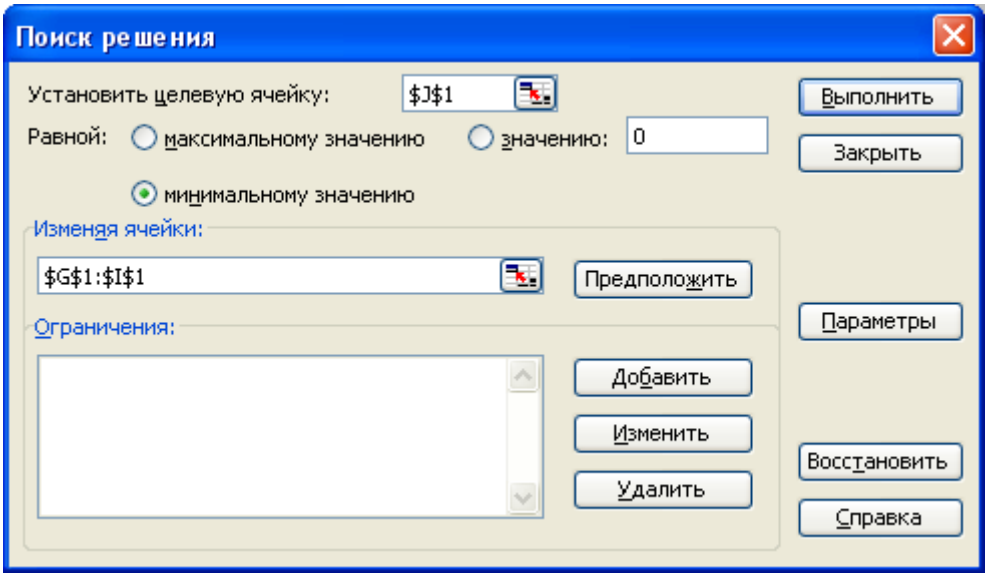


Рисунок 3 — Окно модуля «Поиск решения»

Задача 3. Экспериментально изучалось октановое число бензина, содержащего различные концентрации двух добавок А и В. Пусть Y — октановое число, X_1 — процент первой добавки и X_2 — процент второй добавки.

Предположим, что эффекты добавок А и В складываются, тогда для описания зависимости Y от X_1 и X_2 можно использовать множественную линейную регрессию:

$$y = \beta_0 + \beta_1 x_1 + \beta_2 x_2 + e.$$

Каждая из двух независимых переменных принимала одно из четырех фиксированных значений, а значение Y определялось для каждой комбинации значений $X_1 = x_1$ и $X_2 = x_2$. Анализируемые данные приведены в таблице 3 (файл «Октановое число.xls»).

Таблица 3 – Зависимость октанового числа бензина от добавок

X_1	X_2	Y	X_1	X_2	Y	X_1	X_2	Y	X_1	X_2	Y
2	2	96,3	3	2	95,1	4	2	96,2	5	2	97,8
	3	95,7		3	97,8		3	100,1		3	102,2
	4	99,9		4	99,3		4	103,2		4	104,7
	5	99,4		5	104,9		5	104,3		5	108,8

Найдем коэффициенты регрессии. Для этого в первую очередь найдем сумму квадратов ошибок:

а) откройте файл «Октановое число.xls» и сохраните его как файл «Октановое число 1.xls».

б) введите в ячейку A19 начальное приближение для $b_0 = 90$.

в) введите в ячейку B19 начальное приближение для $b_1 = 1$.

г) введите в ячейку C19 начальное приближение для $b_2 = 1$.

д) рассчитайте в столбце G значения $y_i = b_0 + b_1 x_{1i} + b_2 x_{2i}$ (введите в ячейку D2 формулу «=A$19+B$19*A2+B2*C$19» и скопируйте ее в ячейки D3:D17).

е) рассчитайте в столбце E значения ошибок (остатков) $(y_i - y_i')$ (введите в ячейку E2 формулу «=C2-D2» и скопируйте ее в ячейки E3:E17).

ж) рассчитайте в столбце F значения квадрата ошибок $(y_i - y_i')^2$ (введите в ячейку F2 формулу «=E2^2» и скопируйте ее в ячейки F3:F17).

з) рассчитайте в ячейке F18 сумму квадратов ошибок (введите формулу «=СУММ(F2:F17)»).

Для поиска оценок b_0, b_1, b_2 коэффициентов β_1, β_2, β_3 необходимо минимизировать сумму квадратов ошибки по этим коэффициентам. Для этого:

а) выделите ячейку F17, содержащую сумму квадратов ошибки.

б) выберите команду меню «Сервис ⇒ Поиск решения…».

в) на экране появится диалоговое окно «Поиск решения».

г) проверьте целевую ячейку, она должна быть равна «$F$18».

д) выберите вариант оптимизации — минимизация (выберите «Равной: минимальному значению»).

е) укажите ячейки, значения которых будут изменяться в процессе поиска решения (ячейки $A$19:$C$19, содержащие значения оценок b_0, b_1, b_2).

ж) нажмите кнопку «Выполнить».

з) выберите «Сохранить найденное решение».

и) нажмите кнопку «ОК».

Таким образом, оценка уравнения множественной регрессии есть:

$$y' = 84{,}554 + 1{,}832x_1 + 2{,}683x_2 .$$

Задание. Постройте диаграммы остатков от y', x_1 и x_2. Сделайте выводы.

Построим таблицу дисперсионного анализа. Для этого:

а) рассчитайте в ячейке C18 среднее значение по Y (введите в ячейку формулу «=СРЗНАЧ(C2:C17)»).

б) рассчитайте в столбце G значения $(y_i' - \bar{y})^2$ (введите в ячейку G2 формулу «=(D2-C$18)^2» и скопируйте ее в ячейки G3:G17).

в) рассчитайте в ячейке G18 обусловленную регрессией сумму квадратов SS_D (введите в ячейку формулу «=СУММ(G2:G21)»).

г) введите в ячейку G19 число степеней свободы обусловленной регрессией суммы квадратов $\nu_D = 2$ (количество коэффициентов регрессии минус один).

д) рассчитайте в ячейке F19 число степеней свободы остатков $\nu_R = n - \nu_D - 1 = n - 3$ (введите формулу «=СЧЁТ(C2:C17)-3»).

е) рассчитайте в ячейке G20 обусловленный регрессией средний квадрат $MS_D = SS_D / \nu_D$ (введите формулу «=G18/G19»).

ж) рассчитайте в ячейке F20 средний квадрат отклонения $MS_R = SS_R / \nu_R$ (введите формулу «=F18/F19»).

з) рассчитайте в ячейке G21 F-отношение $F = MS_D / MS_R$ (введите формулу «=G20/F20»).

и) рассчитайте в ячейке F21 коэффициентом детерминации $R^2 = SS_D / SS_T$ (введите формулу «=G18/(G18+F18)»), показывающий долю дисперсии, объясненную регрессией Y по X_1 и X_2.

Проверим гипотезу о том, что множественная линейная регрессия Y по X_1 и X_2. не улучшает оценку Y по сравнению со средним значением Y, т. е. гипотезу H_0 : $\beta_1 = \beta_2 = 0$ против альтернативы: оба коэффициента β_1 и β_2 не равны одновременно нулю. Для этого рассчитайте в ячейке G22 P-значение равное площади области под кривой плотности распределения $F(\nu_D, \nu_R)$ справа от F (введите формулу «=FРАСП(J24;H24;H25».

Поскольку *P*-значение очень мало гипотеза не принимается. Так что октановое число линейно зависит от концентрации, по меньшей мере, одной из добавок A или B.

Проверим гипотезы, о равенстве нулю коэффициентов β_1 и β_2. Для этого:

а) рассчитайте в ячейке A21 коэффициент b_{10} простой линейной регрессии Y от X_1 «=ОТРЕЗОК(C2:C17;A2:A17)».

б) рассчитайте в ячейке B21 коэффициент b_{11} простой линейной регрессии Y от X_1 «=НАКЛОН(C2:C17;A2:A17)».

в) рассчитайте в ячейке A22 коэффициент b_{20} простой линейной регрессии Y от X_2 «=ОТРЕЗОК(C2:C17;B2:B17)».

г) рассчитайте в ячейке B22 коэффициент b_{21} простой линейной регрессии Y от X_2 «=НАКЛОН(C2:C17;B2:B17)».

д) рассчитайте в столбце H значения квадрата ошибок $(y_i - b_{10} - b_{11}x_{1i})^2$ (введите в ячейку H2 формулу «=(C2-A$21-B$21*A2)^2» и скопируйте ее в ячейки H3:H17).

е) рассчитайте в столбце I значения квадрата ошибок $(y_i - b_{20} - b_{21}x_{2i})^2$ (введите в ячейку I2 формулу «=(C2-A$22-B$22*B2)^2» и скопируйте ее в ячейки I3:I17).

ж) рассчитайте в ячейке H18 остаточную сумму квадратов SS'_{R1} (введите формулу «=СУММ(H2:H17)»).

з) рассчитайте в ячейке I18 остаточную сумму квадратов SS'_{R2} (введите формулу «=СУММ(I2:I17)»).

и) рассчитайте в ячейке H19 *F*-отношение $F_1 = (SS'_{R1} - SS_R)/MS_R$ (введите формулу «=(H18-F18)/F20»).

к) рассчитайте в ячейке I19 *F*-отношение $F_2 = (SS'_{R2} - SS_R)/MS_R$ (введите формулу «=(I18-F18)/F20»).

л) рассчитайте в ячейке H20 *P*-значение равное площади области под кривой плотности распределения $F(1, \nu_R)$ справа от F_1 (введите формулу «=FРАСП(H19;1;F19).

м) рассчитайте в ячейке H20 *P*-значение равное площади области под кривой плотности распределения $F(1, \nu_R)$ справа от F_2 (введите формулу «=FРАСП(I19;1;F19).

Поскольку оба *P*-значения очень малы гипотеза $H_0 : \beta_1 = 0$ отвергается, равно как и гипотеза $H_0 : \beta_2 = 0$. Следовательно, X_1 дает значимое улучшение предсказания Y по сравнению с предсказанием, получаемым с помощью регрессии Y только по X_2; соответственно X_2 значимо улучшает предсказание Y по сравнению с предсказанием Y с помощью регрессии Y только по X_1.

Дик Дмитрий Иванович

ОБРАБОТКА РЕЗУЛЬТАТОВ ЭКСПЕРИМЕНТА С ПОМОЩЬЮ MICROSOFT EXCEL

ЧАСТЬ 2

Лабораторный практикум
по дисциплине «Компьютерные технологии в науке и образовании» для студентов квалификации (степени) Магистр направления 190500 (552100)

Редактор Е.А. Устюгова

Подписано к печати	Формат 60×84 1/16	Бумага тип. №1
Печать трафаретная	Усл. печ. л. 1,75	Уч.–изд. л. 1,75
Заказ	Тираж 30	Цена свободная

Редакционно-издательский центр КГУ.
640069, г. Курган, ул. Гоголя, 25.
Курганский государственный университет.