

Множественная регрессия

Спецификация модели

Модель множественной регрессии — это уравнение, отражающее корреляционную связь между результатом и несколькими факторами. В общем виде оно может быть записано в виде:

$$y = f(x_0, \dots, x_m, e),$$

где

- y — зависимая переменная (результат),
- x_0, \dots, x_m — независимые переменные (факторы),
- f — некоторая функция,
- e — случайный остаток.

При проведении регрессионного анализа всегда предполагается, что наблюдения, на основе которых он проводится, были получены по однородной совокупности единиц. Для обеспечения статистической достоверности модели количество наблюдений должно быть в 8–10 раз больше количества параметров при факторах.

При построении модели предполагается, что факторы оказывают влияние на результат, причем влияние одного фактора не зависит от влияния других факторов. Однако, корреляционная связь может существовать и между факторами (т. е. присутствует, т. н. **мультиколлинеарность**). Существование корреляционной связи между факторами может быть выявлено с помощью коэффициентов корреляции $r(x_i; x_j)$ между ними, которые можно записать, например, в виде **корреляционной матрицы**:

$$r_x = \begin{pmatrix} 1 & & & \\ r(x_1; x_2) & 1 & & \\ \vdots & \vdots & \ddots & \vdots \\ r(x_1; x_m) & r(x_2; x_m) & \dots & 1 \end{pmatrix}$$

Наличие мультиколлинеарности можно подтвердить, найдя определитель корреляционной матрицы. Если связь между факторами отсутствует, то определитель равен единице. Если связь между факторами является сильной, то определитель близок к нулю.

Дублирующие факторы с коэффициентом корреляции $|r(x_i; x_j)| \geq 0,7$ исключаются из модели. Предпочтение отдается тому фактору, который наименее связан с другими факторами.

Существует несколько методов спецификации модели:

- **метод последовательного включения факторов** предполагает, что сначала будет построена модель с фактором, наиболее тесно связанным с результатом, а затем поочередно добавляются другие факторы;
- **метод последовательного исключения факторов** предполагает, что сначала строится модель с максимально большим количеством факторов, из которой затем поочередно исключаются незначимые факторы;
- **шаговый регрессионный анализ** является продолжением метода включения. Построение модели начинается с расчета парной регрессии с фактором, наиболее тесно связанным с результатом. А затем добавление каждого нового фактора сопровождается не только оценкой значимости включения данного фактора, но и проверкой значимости факторов уже включенных в модель. Выявленные незначимые факторы исключаются из модели.
- **ступенчатый регрессионный анализ** начинается также с расчета парной регрессии с фактором, наиболее тесно связанным с результатом. Затем вычисляются регрессионные остатки и строится уравнение их зависимости от следующего по степени влияния на результат фактора. По этому уравнению опять вычисляются остатки и процедура повторяется до тех пор пока получаются значимые уравнения

Линейная модель

Говоря о {линейных} эконометрических моделях с несколькими объясняющими переменными, мы фактически исходим о существовании теоретического соотношения

$$y = \theta_0 x_0 + \dots + \theta_m x_m + e$$

между переменными y и x_0, \dots, x_m . Если в правую часть такого соотношения включается константа, то в качестве x_0 тождественно берется единица.

Обращенная к статистическим данным линейная эконометрическая модель с $(m+1)$ -й объясняющей переменной имеет вид:

$$y_i = \theta_0 x_{i0} + \dots + \theta_m x_{im} + e_i, \quad i = 1, \dots, n,$$

где

- y_i — значение результата в i -м наблюдении,
- x_{ij} — значение j -го фактора в i -м наблюдении,
- θ_j — коэффициент при j -м факторе,
- n — количество наблюдений,
- e_i — случайная ошибка в i -м наблюдении.

Запишем линейную модель в матричной форме. Для этого обозначим через

$$\mathbf{y} = (y_1, \dots, y_n)^T, \quad \boldsymbol{\theta} = (\theta_0, \dots, \theta_m)^T, \quad \mathbf{e} = (e_1, \dots, e_n)^T,$$

$$\mathbf{X} = \begin{pmatrix} x_{10} & x_{11} & \dots & x_{1m} \\ x_{20} & x_{21} & \dots & x_{2m} \\ \vdots & \vdots & \ddots & \vdots \\ x_{n0} & x_{n1} & \dots & x_{nm} \end{pmatrix}, \quad \mathbf{X}^T = \begin{pmatrix} x_{10} & x_{20} & \dots & x_{n0} \\ x_{11} & x_{21} & \dots & x_{n1} \\ \vdots & \vdots & \ddots & \vdots \\ x_{1m} & x_{2m} & \dots & x_{nm} \end{pmatrix}.$$

Тогда линейную модель можно записать в матричном виде:

$$\mathbf{y} = \mathbf{X} \cdot \boldsymbol{\theta} + \mathbf{e}$$

Оценивание неизвестных коэффициентов модели можно провести методом наименьших квадратов

$$RSS = \sum_{i=1}^n \left(y_i - \sum_{j=0}^m \theta_j x_{ij} \right)^2 = (\mathbf{y} - \mathbf{X}\boldsymbol{\theta})^T (\mathbf{y} - \mathbf{X}\boldsymbol{\theta}) \rightarrow \min.$$

Приравнивание к нулю частных производных этой суммы по каждой из переменных $\theta_0, \dots, \theta_m$ приводит к системе нормальных уравнений, которую можно записать в матричном виде

$$\mathbf{X}^T \mathbf{X} \boldsymbol{\theta} = \mathbf{X}^T \mathbf{y}.$$

Система имеет единственное решение, которое указывает на точку минимума, если выполнено условие **идентификации**

$$\det(\mathbf{X}^T \mathbf{X}) \neq 0.$$

Решение системы нормальных уравнений имеет вид:

$$\tilde{\boldsymbol{\theta}} = (\mathbf{X}^T \mathbf{X})^{-1} \mathbf{X}^T \mathbf{y}.$$

Обозначая через

$$\tilde{\mathbf{y}} = (\tilde{y}_1, \dots, \tilde{y}_n)^T$$

вектор прогнозных значений результата, получим

$$\tilde{\mathbf{y}} = \mathbf{X}\tilde{\boldsymbol{\theta}} = \mathbf{X}(\mathbf{X}^T \mathbf{X})^{-1} \mathbf{X}^T \mathbf{y}.$$

Вектор остатков \mathbf{e} можно записать

$$\mathbf{e} = \mathbf{y} - \tilde{\mathbf{y}} = \mathbf{y} - \mathbf{X}(\mathbf{X}^T \mathbf{X})^{-1} \mathbf{X}^T \mathbf{y} = (\mathbb{I}_n - \mathbf{X}(\mathbf{X}^T \mathbf{X})^{-1} \mathbf{X}^T) \mathbf{y}.$$

Отметим, что

$$\begin{aligned} \tilde{\mathbf{y}}^T \mathbf{e} &= \tilde{\mathbf{y}}^T (\mathbf{y} - \tilde{\mathbf{y}}) = \tilde{\mathbf{y}}^T \mathbf{y} - \tilde{\mathbf{y}}^T \tilde{\mathbf{y}} = \\ &= \mathbf{y}^T \mathbf{X}(\mathbf{X}^T \mathbf{X})^{-1} \mathbf{X}^T \mathbf{y} - \mathbf{y}^T \mathbf{X}(\mathbf{X}^T \mathbf{X})^{-1} \mathbf{X}^T \mathbf{X}(\mathbf{X}^T \mathbf{X})^{-1} \mathbf{X}^T \mathbf{y} = \mathbf{0}, \\ \mathbf{X}^T \mathbf{e} &= \mathbf{X}^T (\mathbf{y} - \mathbf{X}(\mathbf{X}^T \mathbf{X})^{-1} \mathbf{X}^T \mathbf{y}) = \mathbf{X}^T \mathbf{y} - \mathbf{X}^T \mathbf{X}(\mathbf{X}^T \mathbf{X})^{-1} \mathbf{X}^T \mathbf{y} = \mathbf{0}. \end{aligned}$$

Для остаточной суммы квадратов получаем

$$\begin{aligned} RSS &= |\mathbf{e}|^2 = \mathbf{e}^T \mathbf{e} = (\mathbf{y} - \mathbf{X}\tilde{\boldsymbol{\theta}})^T (\mathbf{y} - \mathbf{X}\tilde{\boldsymbol{\theta}}) = \\ &= \mathbf{y}^T \mathbf{y} - \tilde{\boldsymbol{\theta}}^T \mathbf{X}^T \mathbf{y} - \mathbf{y}^T \mathbf{X}\tilde{\boldsymbol{\theta}} + \tilde{\boldsymbol{\theta}}^T \mathbf{X}^T \mathbf{X}\tilde{\boldsymbol{\theta}} = \\ & \text{(поскольку } \mathbf{y}^T \mathbf{X}\tilde{\boldsymbol{\theta}} \text{ — скаляр, то } \mathbf{y}^T \mathbf{X}\tilde{\boldsymbol{\theta}} = \tilde{\boldsymbol{\theta}}^T \mathbf{X}^T \mathbf{y}) \\ &= |\mathbf{y}|^2 - \tilde{\boldsymbol{\theta}}^T \mathbf{X}^T \mathbf{y} - \tilde{\boldsymbol{\theta}}^T \mathbf{X}^T \mathbf{y} + \tilde{\boldsymbol{\theta}}^T \mathbf{X}^T \mathbf{X}\tilde{\boldsymbol{\theta}} = \\ &= |\mathbf{y}|^2 - 2\tilde{\boldsymbol{\theta}}^T \mathbf{X}^T \mathbf{y} + \tilde{\boldsymbol{\theta}}^T (\mathbf{X}^T \mathbf{X}\tilde{\boldsymbol{\theta}} - \mathbf{X}^T \mathbf{y}). \end{aligned}$$

Из системы нормальных уравнений

$$\mathbf{X}^T \mathbf{X}\tilde{\boldsymbol{\theta}} = \mathbf{X}^T \mathbf{y}$$

имеем

$$\begin{aligned} RSS &= |\mathbf{e}|^2 = |\mathbf{y}|^2 - 2\tilde{\boldsymbol{\theta}}^T \mathbf{X}^T \mathbf{y} = |\mathbf{y}|^2 - 2\tilde{\mathbf{y}}^T \mathbf{y} = |\mathbf{y}|^2 - \tilde{\mathbf{y}}^T (\tilde{\mathbf{y}} + \mathbf{e}) = \\ &= |\mathbf{y}|^2 - |\tilde{\mathbf{y}}|^2 - \tilde{\mathbf{y}}^T \mathbf{e} = |\mathbf{y}|^2 - |\tilde{\mathbf{y}}|^2 \end{aligned}$$

теорему Пифагора в \mathbb{R}^n .

$$|\mathbf{y}|^2 = |\tilde{\mathbf{y}}|^2 + |\mathbf{e}|^2.$$

Уравнение регрессии, построенное по исходным данным, называется моделью **в натуральной форме**. Если же провести предварительную стандартизацию переменных, входящих в модель, т. е. вычесть среднее и поделить на корень из дисперсии

$$y^* = \frac{y - \bar{y}}{\sqrt{S_y^2}}, \quad x_k^* = \frac{x_k - \bar{x}_k}{\sqrt{S_{x_k}^2}}, \quad k = 1, \dots, m,$$

а затем построить по новым переменным модель множественной линейной регрессии

$$\widetilde{y}^* = \beta_1 x_1^* + \dots + \beta_m x_m^*$$

то такая модель будет называться **моделью в стандартизированной форме**.

Применение метода наименьших квадратов к стандартизированной форме модели дает следующую систему нормальных уравнений

$$\begin{cases} \beta_1 & + & \beta_2 r(x_1; x_2) & + & \dots & + & \beta_m r(x_1; x_m) & = & r(x_1; y) \\ \beta_1 r(x_2; x_1) & + & \beta_2 & + & \dots & + & \beta_m r(x_2; x_m) & = & r(x_2; y) \\ \vdots & \vdots & \vdots & \vdots & \ddots & \vdots & \vdots & \vdots & \vdots \\ \beta_1 r(x_m; x_1) & + & \beta_2 r(x_m; x_2) & + & \dots & + & \beta_m & = & r(x_m; y) \end{cases}$$

Последняя система уравнений может быть записана в матричном виде

$$\mathbf{r}_x \boldsymbol{\beta} = \mathbf{r}_{yx},$$

где

$$\bullet \quad \mathbf{r}_x = \begin{pmatrix} 1 & & & \\ r(x_1; x_2) & 1 & & \\ \vdots & \vdots & \ddots & \vdots \\ r(x_1; x_m) & r(x_2; x_m) & \dots & 1 \end{pmatrix}$$

— матрица межфакторной корреляции,

- $\boldsymbol{\beta} = (\beta_1, \dots, \beta_m)^T$ — вектор стандартизированных коэффициентов,
- $\mathbf{r}_{yx} = (r(x_1; y), \dots, r(x_m; y))^T$ — вектор коэффициентов корреляции между результатом y и факторами.

Решение системы имеет вид:

$$\boldsymbol{\beta} = (\mathbf{r}_x)^{-1} \cdot \mathbf{r}_{yx}.$$

Зная коэффициенты β_j стандартизированного уравнения можно найти коэффициенты «чистой» регрессии θ_j :

$$\theta_0 = \bar{y} - \theta_1 \bar{x}_1 - \dots - \theta_m \bar{x}_m, \quad \theta_j = \beta_j \sqrt{\frac{S_y^2}{S_{x_j}^2}}, \quad j = 1, \dots, m.$$

Отсюда, стандартизированные коэффициенты регрессии β_j показывают на сколько среднеквадратических отклонений S_y в среднем изменится результат y при изменении фактора x_j на одно среднеквадратичное отклонение S_{x_j} при фиксированном уровне других факторов, включенных в модель.

Нормальная линейная модель

Будем предполагать следующее:

1. Модель наблюдений результата y с $(m+1)$ фактором x_0, \dots, x_m имеет вид:

$$y_i = \theta_0 x_{i0} + \dots + \theta_m x_{im} + e_i, \quad i = 1, \dots, n,$$

где

y_i — значение результата в i -м наблюдении,

x_{ij} — известное фиксированное значение j -го фактора в i -м наблюдении,

θ_j — неизвестный коэффициент при j -м факторе,

e_i — случайная ошибка в i -м наблюдении.

В матричном виде модель имеет вид:

$$\mathbf{y} = \mathbf{X} \cdot \boldsymbol{\theta} + \mathbf{e}.$$

2. Случайные величины e_1, \dots, e_n независимы в совокупности, имеют нормальное распределение $\mathcal{N}(0; \sigma^2)$ с нулевым математическим ожиданием и дисперсией $\sigma^2 > 0$. Или, иначе, вектор $\mathbf{e} = (e_1, \dots, e_n)$ имеет n -мерное нормальное распределение с математическим ожиданием равным нулевому вектору $(0, \dots, 0)^T$ и диагональной ковариационной матрицей $\text{Cov}(\mathbf{e}) = \sigma^2 \mathbb{I}_n$, где \mathbb{I}_n — единичная матрица размера $n \times n$.

В дальнейшем на предположения этого пункта будем ссылаться как на **стандартные предположения об ошибках** в линейной модели наблюдений.

3. Если не оговорено противное, то в число объясняющих переменных включается переменная, тождественно равная единице

$$x_{i0} = 1, \quad i = 1, \dots, n.$$

4. Определитель матрицы $\mathbf{X}^T \mathbf{X}$ отличен от нуля:

$$\det(\mathbf{X}^T \mathbf{X}) \neq 0,$$

что можно заменить условием: столбцы матрицы \mathbf{X} линейно независимы.

Свойства оценок коэффициентов

При сделанных стандартных предположениях об ошибках модели величины y_1, \dots, y_n являются независимыми в совокупности нормально распределенными случайными величинами

$$y_i \sim \mathcal{N}(\theta_0 x_{i0} + \dots + \theta_m x_{im}; \sigma^2),$$

для которых математические ожидания и дисперсии равны соответственно:

$$\mathbb{E}(y_i) = \theta_0 x_{i0} + \dots + \theta_m x_{im}, \quad \mathbb{D}(y_i) = \sigma^2.$$

Или в матричном виде

$$\mathbb{E}(\mathbf{y}) = \mathbf{X}\boldsymbol{\theta}, \quad \text{Cov}(\mathbf{y}) = \sigma^2 \mathbb{I}_n.$$

Обозначим матрицу $(\mathbf{X}^T \mathbf{X})^{-1} \mathbf{X}^T$ из представления случайного вектора $\tilde{\boldsymbol{\theta}}$ через \mathbf{C} :

$$\tilde{\boldsymbol{\theta}} = (\mathbf{X}^T \mathbf{X})^{-1} \mathbf{X}^T \mathbf{y} = \mathbf{C} \mathbf{y}.$$

Тогда величина $\tilde{\boldsymbol{\theta}}$ является линейным преобразованием нормально распределенного случайного вектора \mathbf{y} и, следовательно, имеет нормальное распределение. Математическое ожидание этого случайного вектора равно:

$$\mathbb{E}\tilde{\boldsymbol{\theta}} = \mathbb{E}((\mathbf{X}^T \mathbf{X})^{-1} \mathbf{X}^T \mathbf{y}) = (\mathbf{X}^T \mathbf{X})^{-1} \mathbf{X}^T \mathbb{E} \mathbf{y} = (\mathbf{X}^T \mathbf{X})^{-1} \mathbf{X}^T \mathbf{X} \boldsymbol{\theta} = \boldsymbol{\theta}.$$

Ковариационная матрица вектора $\tilde{\theta}$ равна

$$\begin{aligned}\text{Cov}(\tilde{\theta}) &= \text{Cov}(\mathbf{C}\mathbf{y}) = \mathbf{C} \text{Cov}(\mathbf{y}) \mathbf{C}^T = \\ &= (\mathbf{X}^T \mathbf{X})^{-1} \mathbf{X}^T \sigma^2 \mathbb{I}_n ((\mathbf{X}^T \mathbf{X})^{-1} \mathbf{X}^T)^T = \\ &= \sigma^2 (\mathbf{X}^T \mathbf{X})^{-1} \mathbf{X}^T \mathbf{X} (\mathbf{X}^T \mathbf{X})^{-1} = \sigma^2 (\mathbf{X}^T \mathbf{X})^{-1}.\end{aligned}$$

Отсюда, в частности, получаем выражение для дисперсии

$$\mathbb{D}(\tilde{\theta}_j) = \sigma^2 (\mathbf{X}^T \mathbf{X})_{jj}^{-1},$$

где $(\mathbf{X}^T \mathbf{X})_{jj}^{-1}$ — j -й диагональный элемент матрицы $(\mathbf{X}^T \mathbf{X})^{-1}$.

Для линейных моделей справедлива следующая важная теорема.

Теорема (Гаусса – Маркова). Пусть модель наблюдений имеет вид:

$$\mathbf{y} = \mathbf{X}\boldsymbol{\theta} + \mathbf{e}, \quad \mathbb{E}(\mathbf{e}) = \mathbf{0}, \quad \text{Cov}(\mathbf{e}) = \sigma^2 \mathbb{I}_n,$$

где матрица \mathbf{X} имеет линейно независимые столбцы, т. е.

$$\det(\mathbf{X}^T \mathbf{X}) \neq 0,$$

Тогда оценка наименьших квадратов $\tilde{\theta} = (\mathbf{X}^T \mathbf{X})^{-1} \mathbf{X}^T \mathbf{y}$ неизвестного вектора коэффициентов $\boldsymbol{\theta}$ является **наилучшей линейной несмещенной оценкой**, т. е. является **эффективной**.

Если к условия теоремы добавить предположение о нормальности случайных ошибок \mathbf{e} , то оценка $\tilde{\theta}$ является наилучшей среди всех несмещенных оценок, а не только в классе линейных.

Ранее рассматривая линейную модель

$$\mathbf{y} = \mathbf{X}\boldsymbol{\theta} + \mathbf{e}$$

с $\mathbf{e} \sim \mathcal{N}(0; \sigma^2 \mathbb{I}_n)$ установили, что оценка наименьших квадратов $\tilde{\theta}_j$ коэффициента θ_j имеет нормальное распределение с параметрами:

$$\mathbb{E}\tilde{\theta}_j = \theta_j, \quad \mathbb{D}\tilde{\theta}_j = \sigma^2 (\mathbf{X}^T \mathbf{X})_{jj}^{-1}.$$

Поэтому случайная величина

$$\frac{\tilde{\theta}_j - \theta_j}{\sqrt{\sigma^2 (\mathbf{X}^T \mathbf{X})_{jj}^{-1}}}$$

имеет стандартное нормальное распределение.

Воспользоваться последней статистикой для построения доверительного интервала невозможно, т. к. дисперсия σ^2 не известна. Заменим дисперсию σ^2 ее оценкой

$$RMS = \frac{RSS}{n - m - 1}.$$

При выполнении стандартных предположений о модели величина

$$\frac{RSS}{\sigma^2}$$

имеет распределение χ^2 с $(n - m - 1)$ степенями свободы. Поэтому RMS несмещенная оценка параметра σ^2 , а случайная величина

$$\frac{\tilde{\theta}_j - \theta_j}{\sqrt{\frac{RSS}{n - m - 1} (\mathbf{X}^T \mathbf{X})_{jj}^{-1}}}$$

имеет распределение Стьюдента с $(n - m - 1)$ степенями свободы.

С доверительной вероятностью $\beta = 1 - \alpha$ интервальной оценкой коэффициента θ_j является

$$\theta_j \geq \tilde{\theta}_j - t_{1-\alpha/2}[n - m - 1] \sqrt{\frac{RSS}{n - m - 1} (\mathbf{X}^T \mathbf{X})_{jj}^{-1}},$$

$$\theta_j \leq \tilde{\theta}_j + t_{1-\alpha/2}[n - m - 1] \sqrt{\frac{RSS}{n - m - 1} (\mathbf{X}^T \mathbf{X})_{jj}^{-1}}$$

или

$$\theta_j \geq \tilde{\theta}_j - t_{(1+\beta)/2}[n - m - 1] \sqrt{\frac{RSS}{n - m - 1} (\mathbf{X}^T \mathbf{X})_{jj}^{-1}},$$

$$\theta_j \leq \tilde{\theta}_j + t_{(1+\beta)/2}[n - m - 1] \sqrt{\frac{RSS}{n - m - 1} (\mathbf{X}^T \mathbf{X})_{jj}^{-1}},$$

где $t_p[k]$ — квантиль распределения Стьюдента с k степенями свободы уровня p .

Множественная корреляция

В качестве качества множественной регрессии может выступать **множественный индекс детерминации** R^2 и **нормированный индекс детерминации** R_{adj}^2 (adjusted R^2):

$$R^2 = 1 - \frac{RSS}{TSS}, \quad R_{adj}^2 = 1 - \frac{\frac{RSS}{n-m-1}}{\frac{TSS}{n-1}},$$
$$R_{adj}^2 = 1 - (1 - R^2) \frac{n-1}{n-m-1},$$

где

$$RSS = \sum_{i=1}^n (y_i - \tilde{y}_i)^2, \quad TSS = \sum_{i=1}^n (y_i - \bar{y})^2,$$
$$ESS = TSS - RSS = \sum_{i=1}^n (\tilde{y}_i - \bar{y})^2.$$

Теорема. Для линейной модели множественный индекс детерминации R^2 выражается через скорректированные коэффициенты регрессии и парные коэффициенты корреляции:

$$R^2 = \sum_{k=1}^m \beta_k r(y; x_k).$$

Доказательство. Уравнение регрессии в стандартизированной форме имеет вид:

$$\tilde{y}^* = \sum_{k=1}^m \beta_k x_k^*.$$

Запишем индекс детерминации в стандартизированном виде:

$$\begin{aligned} R^2 &= 1 - \frac{\sum_{i=1}^n (y_i - \tilde{y}_i)^2}{\sum_{i=1}^n (y_i - \bar{y})^2} = \frac{\sum_{i=1}^n (\tilde{y}_i - \bar{y})^2}{\sum_{i=1}^n (y_i - \bar{y})^2} = \\ &= \frac{\sum_{i=1}^n (\tilde{y}_i^*)^2}{\sum_{i=1}^n (y_i^*)^2} = \frac{\sum_{i=1}^n (\tilde{y}_i^*)^2}{n} = \\ &= \frac{1}{n} \sum_{i=1}^n \tilde{y}_i^* \left(\sum_{k=1}^m \beta_k x_{ik}^* \right) = \sum_{k=1}^m \beta_k \left(\frac{1}{n} \sum_{i=1}^n \tilde{y}_i^* x_{ik}^* \right) = \sum_{k=1}^m \beta_k r(\tilde{y}; x_k). \end{aligned}$$

Преобразуем

$$\begin{aligned}
 R^2 &= \sum_{k=1}^m \beta_k r(\tilde{y}; x_k) = \sum_{k=1}^m \beta_k r(y - e; x_k) = \\
 &= \sum_{k=1}^m \beta_k r(y; x_k) - \sum_{k=1}^m \beta_k r(e; x_k) = \sum_{k=1}^m \beta_k r(y; x_k) - 0 = \sum_{k=1}^m \beta_k r(y; x_k),
 \end{aligned}$$

т. к.

$$r(e; x_k) = \frac{1}{n} \boldsymbol{\delta}_k \mathbf{X}^T \mathbf{e} = \frac{1}{n} \boldsymbol{\delta}_k \mathbf{0}^T = 0,$$

где $\boldsymbol{\delta}_k$ — вектор размера $1 \times n$, i -я компонента которого равна

$$\delta_{ik} = \begin{cases} 1 & \text{при } i = k, \\ 0 & \text{при } i \neq k. \end{cases}$$

Теорема доказана.

Теорема. Для линейной модели множественный индекс детерминации R^2 выражается через парные коэффициенты корреляции:

$$R^2 = 1 - \frac{\det(\mathbf{r}_{yxx})}{\det(\mathbf{r}_x)}.$$

где \mathbf{r}_x — корреляционная матрица факторов x_1, \dots, x_m , \mathbf{r}_{yxx} — корреляционная матрица величин y, x_1, \dots, x_m .

Доказательство. Ранее показали, что стандартизированные коэффициенты β_k являются решениями системы

$$\begin{cases} \beta_1 & + & \beta_2 r(x_1; x_2) & + & \dots & + & \beta_m r(x_1; x_m) & = & r(x_1; y) \\ \beta_1 r(x_2; x_1) & + & \beta_2 & + & \dots & + & \beta_m r(x_2; x_m) & = & r(x_2; y) \\ \vdots & \vdots & \vdots & \vdots & \ddots & \vdots & \vdots & \vdots & \vdots \\ \beta_1 r(x_m; x_1) & + & \beta_2 r(x_m; x_2) & + & \dots & + & \beta_m & = & r(x_m; y) \end{cases}$$

Решая систему методом Крамера, получаем:

$$\beta_k = \frac{\Delta_k}{\det(\mathbf{r}_x)},$$

где определитель Δ_k получается из определителя $\det(\mathbf{r}_x)$ заменой k -го столбца на столбец $(r(x_1; y), \dots, r(x_m; y))^T$.

Преобразуем величину из условия:

$$\begin{aligned}
 1 - \frac{\det(\mathbf{r}_{yxx})}{\det(\mathbf{r}_x)} &= \frac{\det(\mathbf{r}_x) - \det(\mathbf{r}_{yxx})}{\det(\mathbf{r}_x)} = \\
 & \text{(разлагая } \det(\mathbf{r}_{yxx}) \text{ по первой строке)} \\
 &= \frac{1}{\det(\mathbf{r}_x)} (\det(\mathbf{r}_x) - (\det(\mathbf{r}_x) - r(x_1; y)\Delta_1 - \dots - r(x_m; y)\Delta_m)) = \\
 &= \sum_{k=1}^m r(x_k; y) \frac{\Delta_k}{\det(\mathbf{r}_x)} = \sum_{k=1}^m r(x_k; y)\beta_k = R^2.
 \end{aligned}$$

Теорема доказана.

Частная корреляция

Ранжирование факторов, участвующих в множественной линейной регрессии, может быть проведено с помощью стандартизированных коэффициентов регрессии β .

Эта же цель может быть достигнута с помощью частных коэффициентов корреляции, характеризующих тесноту связи между результатом и соответствующим фактором при устранении влияния других факторов, включенных в уравнение регрессии.

Показатели частной корреляции представляют собой отношение сокращения остаточной дисперсии за счет включения в анализ нового фактора к остаточной дисперсии, имевшей место до введение нового фактора в модель.

Предположим, что в модели имеются результат y и два нетривиальных фактора x_1, x_2 . Тогда величина

$$r_{yx_1|x_2} = \sqrt{\frac{RSS_{yx_2} - RSS_{yx_1x_2}}{RSS_{yx_2}}}$$

называется **частным коэффициентом корреляции по x_1 первого порядка**. Воспользовавшись формулой

$$RSS = TSS(1 - R^2)$$

можно записать

$$r_{yx_1|x_2} = \sqrt{1 - \frac{1 - R_{yx_1x_2}^2}{1 - R_{yx_2}^2}}.$$

Упрощая, имеем

$$\begin{aligned} r_{yx_1|x_2} &= \sqrt{\frac{R_{yx_1x_2}^2 - R_{yx_2}^2}{1 - R_{yx_2}^2}} = \\ &= \sqrt{\frac{1 - \frac{\begin{vmatrix} 1 & r(y; x_1) & r(y; x_2) \\ r(y; x_1) & 1 & r(x_1; x_2) \\ r(y; x_2) & r(x_1; x_2) & 1 \end{vmatrix}}{\begin{vmatrix} 1 & r(x_1; x_2) \\ r(x_1; x_2) & 1 \end{vmatrix}} - r^2(y; x_2)}{1 - r^2(y; x_2)}} = \\ &= \dots = \frac{r(y; x_1) - r(y; x_2) \cdot r(x_1; x_2)}{\sqrt{(1 - r^2(y; x_2)) \cdot (1 - r^2(x_1; x_2))}}. \end{aligned}$$

Используют и частные коэффициенты корреляции более высоких порядков:

$$r_{yx_1|x_2 \dots x_m} = \sqrt{1 - \frac{1 - R_{yx_1 \dots x_m}^2}{1 - R_{yx_2 \dots x_m}^2}},$$

которые можно определить через частные коэффициенты более низких степеней по рекуррентной формуле

$$r_{yx_1|x_2 \dots x_m} = \frac{r_{yx_1|x_2 \dots x_{m-1}} - r_{yx_m|x_2 \dots x_{m-1}} \cdot r_{x_1x_m|x_2 \dots x_{m-1}}}{\sqrt{(1 - r_{yx_m|x_2 \dots x_{m-1}}^2) \cdot (1 - r_{x_1x_m|x_2 \dots x_{m-1}}^2)}}.$$

В частном случае, если стандартизированное уравнение регрессии имеет вид

$$\tilde{y}^* = \beta_1 x_1^* + \beta_2 x_2^*,$$

где коэффициенты β определяются по формулам

$$\beta_1 = \frac{r(y; x_1) - r(y; x_2) \cdot r(x_1; x_2)}{1 - r^2(x_1; x_2)},$$

$$\beta_2 = \frac{r(y; x_2) - r(y; x_1) \cdot r(x_1; x_2)}{1 - r^2(x_1; x_2)},$$

то, сравнивая эти формулы с частными коэффициентами корреляции, имеем

$$r_{yx_1|x_2} = \beta_1 \sqrt{\frac{1 - r^2(x_1; x_2)}{1 - r^2(y; x_2)}}, \quad r_{yx_2|x_1} = \beta_2 \sqrt{\frac{1 - r^2(x_1; x_2)}{1 - r^2(y; x_1)}}.$$

Частные коэффициенты подтверждают ранжирование факторов по стандартизированным коэффициентам регрессии. Также, зная частные коэффициенты корреляции, можно определить индекс детерминации

$$R_{\{yx_1 \dots x_m\}}^2 = 1 - (1 - r_{yx_1}^2) \cdot (1 - r_{yx_2|x_1}^2) \cdot \dots \cdot (1 - r_{yx_m|x_1 \dots x_{m-1}}^2).$$

Оценка надежности множественной регрессии

Для оценки значимости параметров уравнения множественной линейной регрессии используется критерий Стьюдента.

Обозначим **стандартную ошибку** параметра θ_j через m_j :

$$m_j = \sqrt{\frac{RSS}{n - m - 1} (\mathbf{X}^T \mathbf{X})_{jj}^{-1}}$$

При выполнении гипотезы $\{H_0: \theta_j = 0\}$ случайная величина

$$t_j = \frac{\tilde{\theta}_j}{m_j}$$

имеет распределение Стьюдента с $(n - m - 1)$ степенями свободы.

Задаем уровень значимости α . Затем, сравним фактическое значение статистики t_j с квантилью $t_{1-\alpha/2}[n - m - 1]$ распределения Стьюдента.

Отвергаем гипотезу H_0 (считаем параметр θ_j **значимым**) при

$$|t_j| > t_{1-\alpha/2}[n - m - 1].$$

В противном случае (при $|t_j| < t_{1-\alpha/2}[n - m - 1]$) считаем параметр θ_j незначимым и влияние фактора x_j на результат y не существенным.

Значимость дополнительных факторов, включаемых в уравнение регрессии, можно оценить с помощью **частного F -критерия**:

$$F_j = \frac{RSS_1 - RSS_2}{RSS_2} (n - m - 1),$$

где RSS_1 — остаточная сумма квадратов для модели без фактора x_j ,

RSS_2 — остаточная сумма квадратов для модели с фактором x_j .

Разделив числитель и знаменатель на полную сумму квадратов TSS , получим

$$F_j = \frac{R^2_{yx_1 \dots x_j \dots x_m} - R^2_{yx_1 \dots x_{\{j-1\}} x_{\{j+1\}} \dots x_m}}{1 - R^2_{yx_1 \dots x_j \dots x_m}} \cdot (n - m - 1).$$

Можно доказать, что $F_j = t_j^2$.

Величина F_j имеет распределение Фишера числом степеней свободы 1 и $(n - m - 1)$. Для уровня значимости α вычисляем квантиль $F_{1-\alpha}[1; n - m - 1]$ распределения Фишера уровня $1 - \alpha$.

Если $F_j > F_{1-\alpha}[1; n - m - 1]$, то включение фактора x_j в модель оправдано.

Если $F_j < F_{1-\alpha}[1; n - m - 1]$, то включение фактора x_j в модель не оправдано.

Помимо оценки значимости отдельных факторов F -критерий Фишера используется для проверки значимости уравнения регрессии в целом. Случайная величина

$$F = \frac{EMS}{RMS} = \frac{\frac{ESS}{m}}{\frac{RSS}{n - m - 1}} = \frac{R^2}{1 - R^2} \cdot \frac{n - m - 1}{m}$$

имеет распределение Фишера с числом степеней свободы m и $(n-m-1)$. Уравнение регрессии значимо, если фактическое значение F -критерия больше табличного $F_{1-\alpha}[m; n - m - 1]$, где α — уровень значимости.

Прогнозирование в множественной регрессии

Пусть имеем линейную модель $\mathbf{y} = \mathbf{X}\boldsymbol{\theta} + \mathbf{e}$ с $(m + 1)$ фактором, $\mathbb{E}(\mathbf{e}) = 0$, $\text{Cov}(\mathbf{e}) = \sigma^2 \mathbb{I}_n$. Известны значения результата и факторов в n наблюдениях. Требуется дать прогноз результата y в $(n + 1)$ -м наблюдении, если значения факторов в $(n + 1)$ -м наблюдении известны.

Точечный прогноз результата находится по уравнению регрессии:

$$\tilde{y}_{n+1} = \mathbf{X}_{n+1} \tilde{\boldsymbol{\theta}} = \mathbf{X}_{n+1} (\mathbf{X}^T \mathbf{X})^{-1} \mathbf{X}^T \mathbf{y},$$

где

- \tilde{y}_{n+1} — прогнозируемое значение результата y ,
- $\mathbf{X}_{n+1} = (1; x_{n+1,1}, \dots, x_{n+1,m})$ — вектор-строка значений факторов x_1, \dots, x_m , для которых строим прогноз,
- $\tilde{\boldsymbol{\theta}}$ — оценка параметров регрессии методом наименьших квадратов.

Математическое ожидание и дисперсия величины \tilde{y}_{n+1} составляют:

$$\mathbb{E}(\tilde{y}_{n+1}) = \mathbf{X}_{n+1} \boldsymbol{\theta}.$$

$$\mathbb{D}(\tilde{y}_{n+1}) = \text{Cov}(\mathbf{X}_{n+1} \tilde{\boldsymbol{\theta}}) = \mathbf{X}_{n+1} \text{Cov}(\tilde{\boldsymbol{\theta}}) \mathbf{X}_{n+1}^T = \sigma^2 \mathbf{X}_{n+1} (\mathbf{X}^T \mathbf{X})^{-1} \mathbf{X}_{n+1}^T.$$

Вычислим характеристики **ошибки прогноза** $y_{n+1} - \tilde{y}_{n+1}$:

$$\mathbb{E}(y_{n+1} - \tilde{y}_{n+1}) = \mathbf{X}_{n+1} \boldsymbol{\theta} - \mathbf{X}_{n+1} \boldsymbol{\theta} = 0,$$

$$\mathbb{D}(y_{n+1} - \tilde{y}_{n+1}) = \sigma^2 + \sigma^2 \mathbf{X}_{n+1} (\mathbf{X}^T \mathbf{X})^{-1} \mathbf{X}_{n+1}^T = \sigma^2 (1 + \nu),$$

где $\nu = \mathbf{X}_{n+1} (\mathbf{X}^T \mathbf{X})^{-1} \mathbf{X}_{n+1}^T$.

Величина

$$\frac{y_{n+1} - \tilde{y}_{n+1}}{\sqrt{\sigma^2 (1 + \nu)}}$$

имеет стандартное нормальное распределение, а величина

$$\frac{y_{n+1} - \tilde{y}_{n+1}}{\sqrt{\frac{RSS}{n - m - 1} (1 + \nu)}}$$

имеет распределение Стьюдента с $(n - m - 1)$ степенями свободы.

Интервальную оценку для величины y с доверительной вероятностью β можно записать в виде:

$$\tilde{y}_{n+1} \pm t_{(1+\beta)/2}[n-m-1] \cdot m_{\tilde{y}},$$

где

$$m_{\tilde{y}} = m_y \cdot \sqrt{1+v},$$

$$m_y = \sqrt{\frac{RSS}{n-m-1}}, \quad v = \mathbf{X}_{n+1}(\mathbf{X}^T \mathbf{X})^{-1} \mathbf{X}_{n+1}^T.$$

Пример. До эмпирическим данным, приведенным в следующей таблице

| x_1 | x_2 | y |
|-------|-------|-----|
| 1 | 2 | 3 |
| 4 | 7 | 39 |
| 2 | 3 | 11 |
| 3 | 6 | 27 |
| 3 | 4 | 19 |
| 6 | 6 | 48 |
| 3 | 4 | 19 |
| 7 | 6 | 55 |
| 5 | 3 | 23 |
| 7 | 6 | 55 |

с помощью программы Excel определить:

1. Ожидаемые значения и дисперсии величин x_1, x_2 и y ;
2. Корреляционную матрицу величин x_1, x_2 и y ;
3. Стандартизированные коэффициенты регрессии β_1 и β_2 ;
4. Частные коэффициенты корреляции $r_{yx_1|x_2}$, $r_{yx_2|x_1}$;
5. Коэффициенты регрессии $\theta_0, \theta_1, \theta_2$ в натуральном виде;
6. Индекс детерминации R^2 и нормированный индекс детерминации R_{adj}^2 ;
7. Надежность уравнения регрессии и значимость коэффициентов с помощью F -критериев с уровнем значимости $\alpha = 0,05$;
8. Интервальный прогноз параметра θ_2 с доверительной вероятностью $\beta = 0,9$;

9. Точечный прогноз величины y , если $x_1 = 5$, $x_2 = 4$.

10. Интервальный прогноз величины y с доверительной вероятностью $\beta = 0,9$ при $x_1 = 5$, $x_2 = 4$

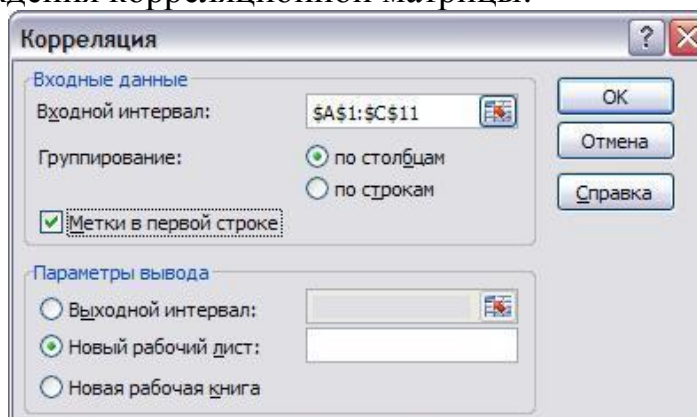
Решение. Запишем исходные данные в MS Excel:

| | A | B | C |
|----|-------|-------|-----|
| 1 | x_1 | x_2 | y |
| 2 | 1 | 2 | 3 |
| C | 4 | 7 | 39 |
| 4 | 2 | 3 | 11 |
| 5 | 3 | 6 | 27 |
| 6 | 3 | 4 | 19 |
| 7 | 6 | 6 | 48 |
| 8 | 3 | 4 | 19 |
| 9 | 7 | 6 | 55 |
| 10 | 5 | 3 | 23 |
| 11 | 7 | 6 | 55 |

1. Объем выборки $n = 10$, количество факторов $m = 2$. Ожидаемое значение величины x_1 можно вычислить с помощью функции «СРЗНАЧ (A2:A11)». Имеем $\bar{x}_1 = 4,1$. Для вычисления дисперсии величины x_1 можно использовать функцию «ДИСПР (A2:A11)». В результате $S_{x_1}^2 = 3,89$. Аналогично,

$$\bar{x}_2 = 4,7, \bar{y} = 29,9, S_{x_2}^2 = 2,61, S_y^2 = 304,49.$$

2. Используем надстройку «Анализ данных» пункт «Корреляция» для нахождения корреляционной матрицы:



В результате получим

| | x1 | x2 | y |
|----|--------|--------|---|
| x1 | 1 | | |
| x2 | 0,6371 | 1 | |
| y | 0,9388 | 0,8432 | 1 |

Заполняем пустые клетки таблицы до симметричной матрицы:

| | x1 | x2 | y |
|----|--------|--------|--------|
| x1 | 1 | 0,6371 | 0,9388 |
| x2 | 0,6371 | 1 | 0,8432 |
| y | 0,9388 | 0,8432 | 1 |

3. Находим обратную матрицу для корреляционной матрицы r_x факторов x_1 и x_2

| | |
|--------|--------|
| 1 | 0,6371 |
| 0,6371 | 1 |

с помощью функции «МОБР (B2 : C3) ». Имеем

| | |
|---------|---------|
| 1,6832 | -1,0723 |
| -1,0723 | 1,6832 |

Умножаем последнюю матрицу $(r_x)^{-1}$ на матрицу r_{yx}

| |
|--------|
| 0,9388 |
| 0,8432 |

с помощью функции «МУМНОЖ». Имеем

| |
|-------|
| 0,676 |
| 0,413 |

Это и есть стандартизированные коэффициенты регрессии

$$\beta_1 = 0,676, \quad \beta_2 = 0,413.$$

4. Частные коэффициенты корреляции $r_{yx_1|x_2}$, $r_{yx_2|x_1}$ вычисляем по формулам

$$r_{yx_1|x_2} = \frac{r(y; x_1) - r(y; x_2) \cdot r(x_1; x_2)}{\sqrt{(1 - r^2(y; x_2)) \cdot (1 - r^2(x_1; x_2))}},$$

$$r_{yx_2|x_1} = \frac{r(y; x_2) - r(y; x_1) \cdot r(x_1; x_2)}{\sqrt{(1 - r^2(y; x_1)) \cdot (1 - r^2(x_1; x_2))}}.$$

Получаем

$$r_{y_{x_1}|x_2} = \frac{0,9388 - 0,8432 \cdot 0,6371}{\sqrt{(1 - 0,8432^2) \cdot (1 - 0,6371^2)}} = 0,969,$$

$$r_{y_{x_2}|x_1} = \frac{0,8432 - 0,9388 \cdot 0,6371}{\sqrt{(1 - 0,9388^2) \cdot (1 - 0,6371^2)}} = 0,339.$$

5. Коэффициенты регрессии $\theta_0, \theta_1, \theta_2$ найдем тремя способами.

a) Воспользуемся формулами

$$\theta_0 = \bar{y} - \theta_1 \bar{x}_1 - \theta_2 \bar{x}_2, \quad \theta_j = \beta_j \sqrt{\frac{S_y^2}{S_{x_j}^2}}, \quad j = 1, 2$$

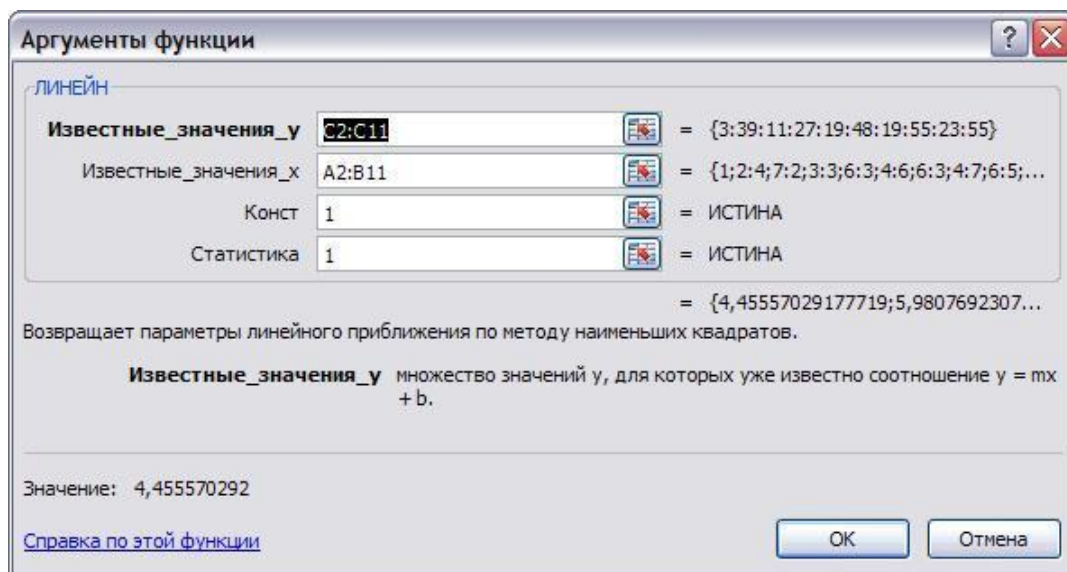
и результатами расчетов пунктов 1 и 3:

$$\theta_1 = 0,675998 * \text{КОРЕНЬ}\left(\frac{304,49}{3,89}\right) = 5,9808,$$

$$\theta_2 = 0,412512 * \text{КОРЕНЬ}\left(\frac{304,49}{2,61}\right) = 4,4556,$$

$$\theta_0 = 29,9 - 5,9808 * 4,1 - 4,4556 * 4,7 = -15,5623.$$

b) Применим функцию «ЛИНЕЙН».



Для этого выделяем 5 строк и 3 столбца (на 1 больше количества факторов x).

| E1 | | fx | | =ЛИНЕЙН(C2:C11;A2:B11;1;1) | | | |
|----|----|----|----|----------------------------|---------|---|---|
| | A | B | C | D | E | F | G |
| 1 | x1 | x2 | y | | 4,45557 | | |
| 2 | 1 | 2 | 3 | | | | |
| 3 | 4 | 7 | 39 | | | | |
| 4 | 2 | 3 | 11 | | | | |
| 5 | 3 | 6 | 27 | | | | |
| 6 | 3 | 4 | 19 | | | | |
| 7 | 6 | 6 | 48 | | | | |
| 8 | 3 | 4 | 19 | | | | |
| 9 | 7 | 6 | 55 | | | | |
| 10 | 5 | 3 | 23 | | | | |
| 11 | 7 | 6 | 55 | | | | |
| 12 | | | | | | | |

Нажимаем клавишу «F2». А затем, удерживая нажатыми «Ctrl» и «Shift», нажимаем «Enter». Получим

| | | |
|----------|---------|----------|
| 4,4556 | 5,9808 | −15,5623 |
| 0,7016 | 0,5747 | 2,6996 |
| 0,9825 | 2,7626 | #Н/Д |
| 195,9809 | 7 | #Н/Д |
| 2991,476 | 53,4244 | #Н/Д |

В таблице содержится следующая информация

| | | |
|------------|-------------|------------|
| θ_2 | θ_1 | θ_0 |
| m_2 | m_1 | m_0 |
| R^2 | m_y | #Н/Д |
| F | $n - m - 1$ | #Н/Д |
| ESS | RSS | #Н/Д |

В первой строке записаны коэффициенты регрессии, во второй — стандартные ошибки коэффициентов, в третьей — индекс детерминации и стандартная ошибка величины y , в четвертой — критерий Фишера и число степеней свободы, в пятой — суммы квадратов, объясненная регрессией и остаточная.

- с) Используем надстройку «Анализ данных» пункт «Регрессия»

В результате получим новый лист

ВЫВОД ИТОГОВ

Регрессионная статистика

| | |
|-------------------------|---------|
| Множественный R | 0,9912 |
| R-квадрат | 0,9825 |
| Нормированный R-квадрат | 0,9774 |
| Стандартная ошибка | 2,7626 |
| Наблюдения | 10,0000 |

Дисперсионный анализ

| | <i>df</i> | <i>SS</i> | <i>MS</i> | <i>F</i> | <i>Значимость F</i> |
|-----------|-----------|-----------|-----------|----------|---------------------|
| Регрессия | 2,0000 | 2991,4756 | 1495,7378 | 195,9809 | 0,0000 |
| Остаток | 7,0000 | 53,4244 | 7,6321 | | |
| Итого | 9,0000 | 3044,9000 | | | |

| | Кэф- фици- енты | Стан- дарт- ная ошибка | t-ста- тис- тика | P- Знач ение | Нижние 95% | Верхние 95% | Нижние 90,0% | Верхние 90,0% |
|--------------------|-----------------------|---------------------------------|------------------------|--------------------|---------------|----------------|-----------------|------------------|
| Y-пересе- чение | -15,562 | 2,6996 | -5,7647 | 0,0007 | -21,9459 | -9,1788 | -20,6770 | -10,4477 |
| x1 | 5,9808 | 0,5747 | 10,4075 | 0,0000 | 4,6219 | 7,3396 | 4,8920 | 7,0695 |
| x2 | 4,4556 | 0,7016 | 6,3509 | 0,0004 | 2,7966 | 6,1145 | 3,1264 | 5,7847 |

Интересующие нас коэффициенты находятся во втором столбце третьей таблицы.

6. Индекс детерминации R^2 и нормированный индекс детерминации R^2_{adj} находятся во второй и третьей строке таблицы «Регрессионная статистика» из надстройки «Анализ данных» пункт «Регрессия». Из таблиц предыдущего пункта находим

$$R^2 = 0,9825, \quad R^2_{adj} = 0,9774.$$

Коэффициент R^2 можно также вычислить, используя функцию «ЛИНЕЙН», а для нахождения R^2_{adj} воспользоваться формулой

$$R^2_{adj} = 1 - (1 - R^2) \frac{n - 1}{n - m - 1}.$$

7. Значимость уравнения регрессии с уровнем значимости $\alpha = 0,05$ проверяется путем сравнения статистики F с квантилью распределения Фишера. Статистику F можно вычислить, используя функцию «ЛИНЕЙН» или надстройку «Анализ данных» пункт «Регрессия»:

$$F = 195,9809.$$

Квантиль распределения Фишера уровня $1 - 0,05 = 0,95$

$$F_{1-0,05}[2; 10 - 2 - 1]$$

можно найти с помощью функции «ФРАСПОВР (0, 05; 2; 7) »

$$F_{0,95}[2; 7] = 4,7374.$$

Так как фактическое значение F -критерия больше табличного $F_{1-\alpha}[m; n - m - 1]$, то уравнение регрессии значимо.

Частные F -критерии вычисляем по формулам:

$$F_1 = \frac{R^2 - r^2(y; x_2)}{1 - R^2} \cdot (10 - 2 - 1) = \frac{0,9825 - 0,8432^2}{1 - 0,9825} \cdot 7 = 108,3162,$$

$$F_2 = \frac{R^2 - r^2(y; x_1)}{1 - R^2} \cdot (10 - 2 - 1) = \frac{0,9825 - 0,9388^2}{1 - 0,9825} \cdot 7 = 40,3345.$$

Также величины F_j можно найти, возведя в квадрат « t -статистики» t_j из третьей таблицы надстройки «Анализ данных. Регрессия»:

$$F_1 = 10,4075^2 = 108,3162, \quad F_2 = 6,3509^2 = 40,3345.$$

Далее, т. к. частные F -критерии больше величины квантили

$$F_{1-0,05}[1; 10 - 2 - 1] = 5,5914.$$

Значит коэффициенты регрессии значимы.

8. а) Границы интервала параметра θ_2 с доверительной вероятностью $\beta = 0,9$ можно вычислить по формулам

$$\theta_2 \geq \tilde{\theta}_2 - t_{(1+0,9)/2}[10 - 2 - 1] \sqrt{\frac{RSS}{10 - 2 - 1} (\mathbf{X}^T \mathbf{X})_{jj}^{-1}},$$

$$\theta_2 \leq \tilde{\theta}_2 + t_{(1+0,9)/2}[10 - 2 - 1] \sqrt{\frac{RSS}{10 - 2 - 1} (\mathbf{X}^T \mathbf{X})_{jj}^{-1}},$$

Для этого вводим в рассмотрение дополнительный фактор x_0 тождественно равный 1.

Матрица \mathbf{X} размером 10 строк и 3 столбца выделена на следующем рисунке желтым цветом. Транспонированную матрицу находим с помощью функции «ТРАНСП (A2 : C11)». Затем выделяем 3 строки и 10 столбцов, нажимаем «F2» и «Ctrl+Shift+Enter».

В результате будет найдена транспонированная матрица X^T , выделенная бирюзовым цветом.

| A13 | | fx {=ТРАНСП(A2:C11)} | | | | | | | | | |
|-----|----|----------------------|----|----|---|---|---|---|---|---|---|
| | A | B | C | D | E | F | G | H | I | J | K |
| 1 | x0 | x1 | x2 | y | | | | | | | |
| 2 | 1 | 1 | 2 | 3 | | | | | | | |
| 3 | 1 | 4 | 7 | 39 | | | | | | | |
| 4 | 1 | 2 | 3 | 11 | | | | | | | |
| 5 | 1 | 3 | 6 | 27 | | | | | | | |
| 6 | 1 | 3 | 4 | 19 | | | | | | | |
| 7 | 1 | 6 | 6 | 48 | | | | | | | |
| 8 | 1 | 3 | 4 | 19 | | | | | | | |
| 9 | 1 | 7 | 6 | 55 | | | | | | | |
| 10 | 1 | 5 | 3 | 23 | | | | | | | |
| 11 | 1 | 7 | 6 | 55 | | | | | | | |
| 12 | | | | | | | | | | | |
| 13 | 1 | 1 | 1 | 1 | 1 | 1 | 1 | 1 | 1 | 1 | |
| 14 | 1 | 4 | 2 | 3 | 3 | 6 | 3 | 7 | 5 | 7 | |
| 15 | 2 | 7 | 3 | 6 | 4 | 6 | 4 | 6 | 3 | 6 | |
| 16 | | | | | | | | | | | |

Произведение матриц $X^T X$ находим с помощью функции «МУМНОЖ (A13:J15;A2:C11)», а обратную матрицу $(X^T X)^{-1}$ с помощью «МОБР (F2:H4)». Нам понадобится элемент из третьей строки и третьего столбца $(X^T X)^{-1}_{33} = 0,064$.

| F6 | | fx {=МОБР(F2:H4)} | | | | | | | | | |
|----|----|-------------------|----|----|---|---------|---------|---------|---|---|---|
| | A | B | C | D | E | F | G | H | I | J | K |
| 1 | x0 | x1 | x2 | y | | | | | | | |
| 2 | 1 | 1 | 2 | 3 | | 10 | 41 | 47 | | | |
| 3 | 1 | 4 | 7 | 39 | | 41 | 207 | 213 | | | |
| 4 | 1 | 2 | 3 | 11 | | 47 | 213 | 247 | | | |
| 5 | 1 | 3 | 6 | 27 | | | | | | | |
| 6 | 1 | 3 | 4 | 19 | | 0,9549 | -0,0192 | -0,1651 | | | |
| 7 | 1 | 6 | 6 | 48 | | -0,0192 | 0,0433 | -0,0337 | | | |
| 8 | 1 | 3 | 4 | 19 | | -0,1651 | -0,0337 | 0,0645 | | | |
| 9 | 1 | 7 | 6 | 55 | | | | | | | |
| 10 | 1 | 5 | 3 | 23 | | | | | | | |
| 11 | 1 | 7 | 6 | 55 | | | | | | | |
| 12 | 1 | 5 | 4 | | | | | | | | |
| 13 | 1 | 1 | 1 | 1 | 1 | 1 | 1 | 1 | 1 | 1 | |
| 14 | 1 | 4 | 2 | 3 | 3 | 6 | 3 | 7 | 5 | 7 | |
| 15 | 2 | 7 | 3 | 6 | 4 | 6 | 4 | 6 | 3 | 6 | |
| 16 | | | | | | | | | | | |

Квантиль распределения Стьюдента $t_{(1+0,9)/2}[10-2-1]$ вычисляем с помощью функции «СТЮДРАСПОБР (1-0,9;7)»

$$t_{0,95}[7] = 1,895.$$

Остальные величины для определения границ доверительного интервала определяются с помощью функции «ЛИНЕЙН»:

| | | |
|----------|---------|----------|
| 4,4556 | 5,9808 | −15,5623 |
| 0,7016 | 0,5747 | 2,6996 |
| 0,9825 | 2,7626 | #Н/Д |
| 195,9809 | 7 | #Н/Д |
| 2991,476 | 53,4244 | #Н/Д |

$$\theta_2 \geq 4,4557 - 1,895 \sqrt{\frac{53,4244}{7} 0,064} = 3,1264,$$

$$\theta_2 \leq 4,4557 + 1,895 \sqrt{\frac{53,4244}{7} 0,064} = 5,7847.$$

b) Интервальный прогноз параметра θ_2 с доверительной вероятностью $\beta = 0,9$ можно также вычислить, используя надстройку «Анализ данных» пункт «Регрессия». Для этого в меню необходимо указать уровень надежности 0,9. Тогда нижнюю и верхнюю границу доверительного интервала можно найти в предпоследнем и последнем столбце третьей таблицы. В данном примере

$$\theta_{2,left} = 3,1264, \quad \theta_{2,right} = 5,7847.$$

9. Точечный прогноз величины y при $x_1 = 5$, $x_2 = 4$ находим по формуле

$$\tilde{y}_{n+1} = \mathbf{X}_{n+1} \tilde{\boldsymbol{\theta}},$$

где $\mathbf{X}_{n+1} = (1; 5; 4)$, $\tilde{\boldsymbol{\theta}} = (-15,5623; 5,9808; 4,4556)^T$. Получаем

$$\tilde{y}_{n+1} = 32,1638.$$

10. Интервальную оценку для величины y с доверительной вероятностью $\beta = 0,9$ можно записать в виде:

$$\tilde{y}_{n+1} \pm t_{(1+0,9)/2} [10 - 2 - 1] \cdot m_{\tilde{y}},$$

где

$$m_{\tilde{y}} = m_y \cdot \sqrt{1 + v},$$

$$m_y = \sqrt{\frac{RSS}{n - m - 1}}, \quad v = \mathbf{X}_{n+1}(\mathbf{X}^T \mathbf{X})^{-1} \mathbf{X}_{n+1}^T.$$

Квантиль $t_{0,95}[7] = 1,895$ вычисляли ранее, стандартная ошибка m_y величины y находится в третьей строке функции «ЛИНЕЙН»:

$$m_y = 2,7626.$$

Обратную матрицу $(\mathbf{X}^T \mathbf{X})^{-1}$ вычислили в пункте 8

$$(\mathbf{X}^T \mathbf{X})^{-1} = \begin{pmatrix} 0,9549 & -0,0192 & -0,1651 \\ -0,0192 & 0,0433 & -0,0337 \\ 0,1651 & -0,0337 & 0,0645 \end{pmatrix}$$

Полагая $\mathbf{X}_{n+1} = (1; 5; 4)$, перемножаем матрицы с помощью функции «МУМНОЖ»:

$$\mathbf{X}_{n+1}(\mathbf{X}^T \mathbf{X})^{-1} = (0,1983; 0,0625; -0,0754).$$

Аналогично, перемножая матрицы, находим:

$$v = \mathbf{X}_{n+1}(\mathbf{X}^T \mathbf{X})^{-1} \mathbf{X}_{n+1}^T = 0,2091.$$

Теперь можем вычислить стандартную ошибку

$$m_{\tilde{y}} = m_y \cdot \sqrt{1 + v} = 2,7626 \cdot \sqrt{1 + 0,2091} = 3,0377$$

и границы 90%-го доверительного интервала

$$\tilde{y}_{left} = 32,1638 - 1,895 \cdot 3,0377 = 26,409,$$

$$\tilde{y}_{right} = 32,1638 + 1,895 \cdot 3,0377 = 37,919.$$