



# Эконометрика.

Четвертая лекция.

# Модель множественной регрессии.

Множественная регрессия – уравнение связи с несколькими независимыми переменными:

Можно записать линейную модель множественной регрессии в двух видах:

$$1. y_i = \beta_1 + \beta_2 x_{i2} + \beta_3 x_{i3} + \dots + \beta_k x_{ik} + U_i$$

$$i = [1; n]$$

$$2. y_i = \beta_1 x_{i1} + \beta_2 x_{i2} + \beta_3 x_{i3} + \dots + \beta_k x_{ik} + U_i$$

если  $x_{i1} = 1$ , для любого  $i = [1; n]$

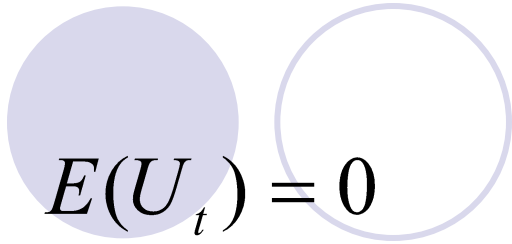
Гипотезы, лежащие в основе множественной модели, являются естественным обобщением модели парной регрессии:

1. Спецификация модели:

$$y_i = \beta_1 x_{i1} + \beta_2 x_{i2} + \dots + \beta_k x_{ik} + U_i \quad \text{для любого } i=[1;n]$$

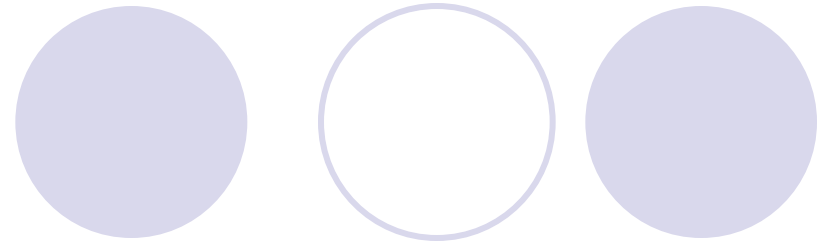
2.  $x_{i1}, x_{i2}, \dots, x_{ik}$  - детерминированные величины  
 $x_S = (x_{1S}, x_{2S}, \dots, x_{nS})^T$  линейно независимо в  $R^n$

3.



$$E(U_t) = 0$$

$$E(U_t^2) = \sigma^2$$



4.

$$E(U_t, U_s) = 0, \forall t \neq s$$

5.

$$U_t \sim N(0, \sigma^2)$$

$$U_t \approx N(0, \sigma^2)$$

Если выполняются эти условия, то модель называется нормальной линейной регрессией.

Введем следующие обозначения:

$$y = (y_1, y_2, \dots, y_n)^T$$

Вектор значений зависимой переменной

$$\beta = (\beta_1, \beta_2, \dots, \beta_k)^T$$

Вектор неизвестных параметров модели

$$U = (U_1, U_2, \dots, U_n)^T$$

Вектор значений случайной компоненты

$$X = \begin{matrix} & x_{11} & x_{12} & x_{13} & \square & x_{1k} \\ x_{21} & x_{22} & x_{22} & \square & x_{2k} \\ \square & \square & \square & \square & \square \\ & x_{n1} & x_{n2} & x_{n3} & \square & x_{nk} \end{matrix}$$

Матрица значений регрессоров



$\hat{\beta}$  – вектор оценок

$$\hat{\beta} = (X^T X)^{-1} X^T * Y$$

# Интерпретация множественного уравнения регрессии.

$$y = \alpha + \beta_1 x_1 + \beta_2 x_2 + \dots + \beta_k x_k + u$$

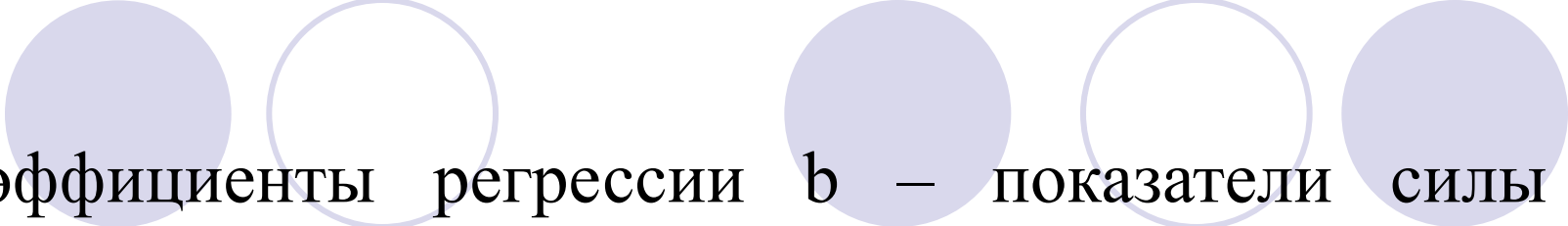
$$\hat{y} = a + b_1 x_1 + b_2 x_2 + \dots + b_k x_k$$

$$\hat{y} = 116,7 + 0,112x_1 - 0,739x_2$$

$x_1$  – доход потребителя (руб.)

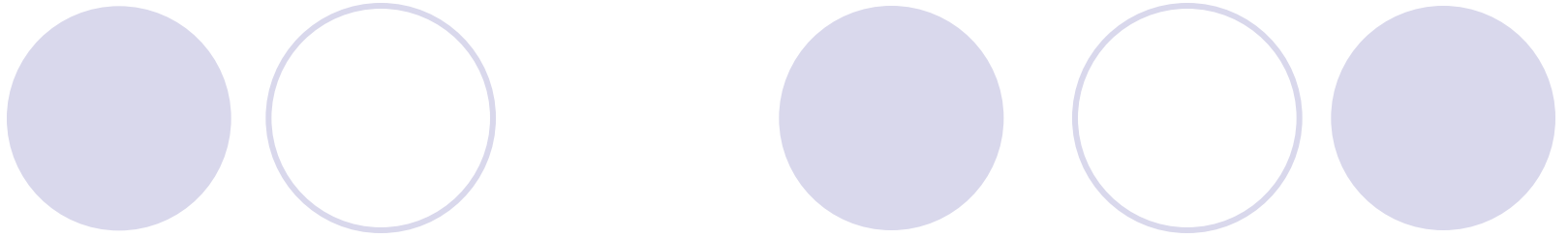
$x_2$  – цена продукта питания (руб.)

$Y$  – расход на питание (руб.)



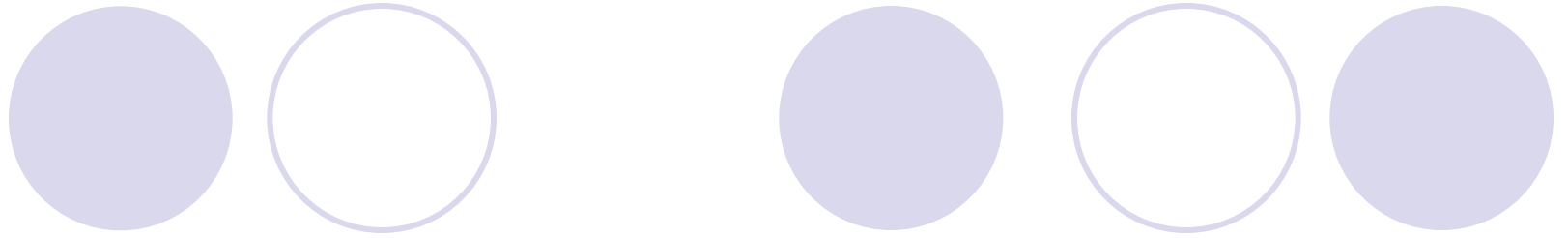
Коэффициенты регрессии  $b$  — показатели силы связи, характеризующие абсолютное изменение результативного признака  $Y$  (в его единицах измерения) при изменении факторного признака  $x$  на 1 единицу своего измерения и при фиксированном влиянии остальных факторов, включенных в модель.





Коэффициент  $\alpha$  показывает совокупное влияние прочих факторов, не включенных в модель.

Используя коэффициенты регрессии можно рассчитать частные коэффициенты эластичности. Как правило их рассчитывают для средних значений факторов:



$$\varepsilon_j = b_j * \frac{\bar{x}_j}{\bar{y}} = b_j * \frac{\bar{x}_j}{a + \sum_{j=1}^k b_j x_j}$$

Частные коэффициенты эластичности имеют тот же смысл, что и обычные, добавляется лишь ограничение на фиксированное значение остальных факторов.

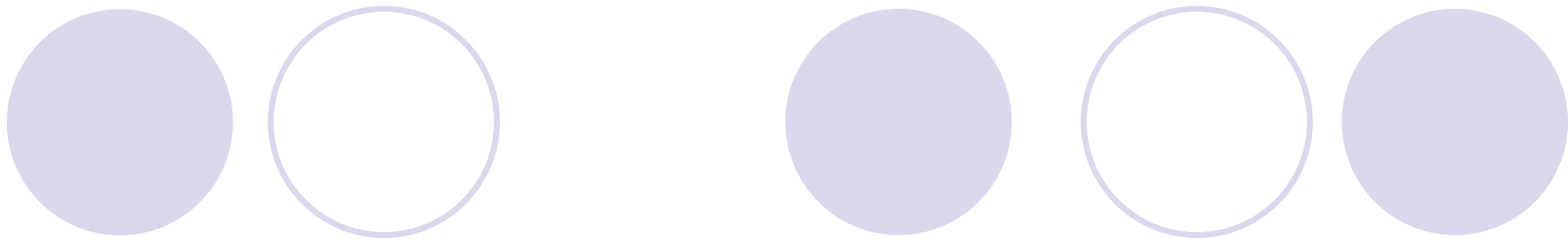


Все коэффициенты регрессии должны быть подвергнуты оценке статистической значимости.

Процедура проверки такая же как и в парной линейной регрессии.



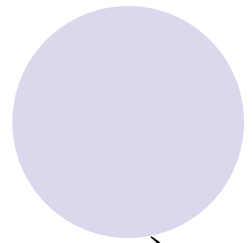
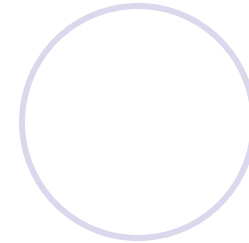
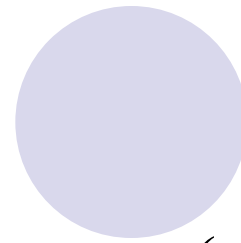
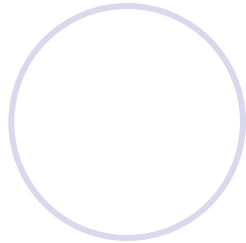
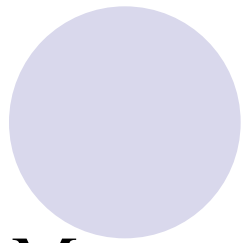
# **Анализ показателей тесноты связи.**



## Парные коэффициенты

$$r_{x_j y} = \frac{\overline{x_j y} - \overline{x_j} \overline{y}}{\sigma_{x_j} \sigma_y}$$

$$r_{x_j x_s} = \frac{\overline{x_j x_s} - \overline{x_j} \overline{x_s}}{\sigma_{x_j} \sigma_{x_s}}$$



Мультиколлинеарность

(коллинеарность)–

ситуация, когда регрессоры тесно связаны между собой. Если объясняющие переменные связаны строгой функциональной зависимостью, то говорят о совершенной мультиколлинеарности.


$$Y = a + b_1 x_1 + b_2 x_2 + b_3 x_3$$

где  $Y$  - общая величина расходов на питание;

$x_1$  - заработная плата;

$x_2$  - доход, получаемый вне работы;

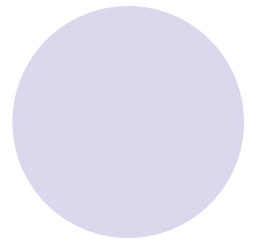
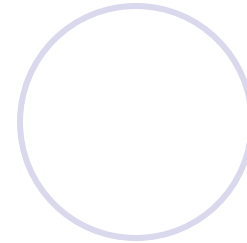
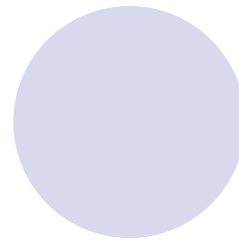
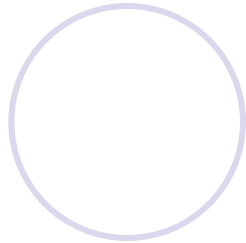
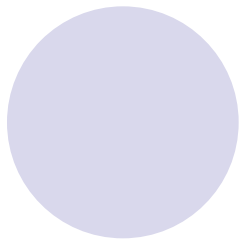
$x_3$  - совокупный доход.



Для оценки мультиколлинеарности составляется и анализируется матрица парных коэффициентов корреляции.

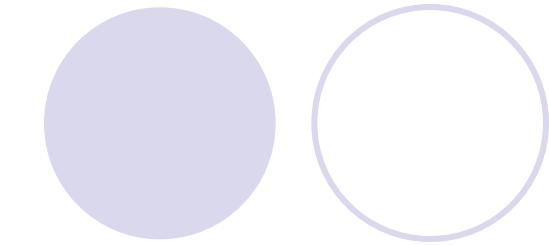
В первой строке и в первом столбце записывают все факторы, начиная с зависимой переменной.

В клетках матрицы рассчитывают соответствующие парные коэффициенты корреляции.



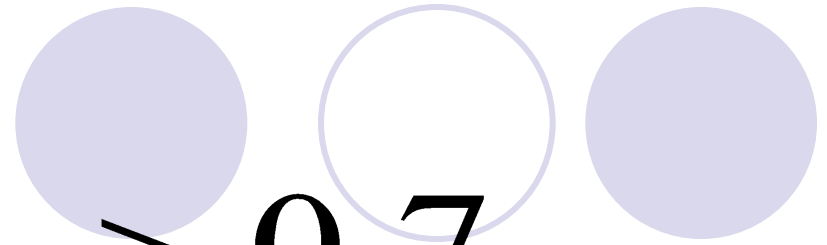
	$Y$	$x_1$	$x_2$	$x_3$
$Y$	1	$r_{yx_1}$	$r_{yx_2}$	$r_{yx_3}$
$x_1$	$r_{x_1y}$	1	$r_{x_1x_2}$	$r_{x_1x_3}$
$x_2$	$r_{x_2y}$	$r_{x_2x_1}$	1	$r_{x_2x_3}$
$x_3$	$r_{x_3y}$	$r_{x_3x_1}$	$r_{x_3x_2}$	1





Если

$$r_{x_j x_s} > 0,7$$



тогда считают, что регрессоры коллинеарны.

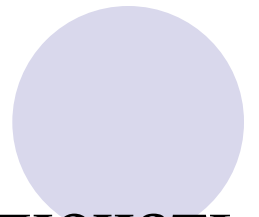
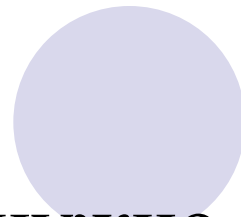
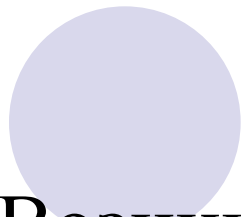
Т.е. между регрессорами существует тесная

связь. В этом случае нельзя определить их

изолированное влияние на результативный

показатель и параметры уравнения регрессии

оказываются *неинтерпретируемыми*.



Возникает вопрос: нужно ли исключать коррелируемые регрессоры?

Однозначного ответа на этот вопрос нет. Существует даже такая школа, представители которой считают, что и не нужно ничего делать, поскольку «так устроен мир».



Другие

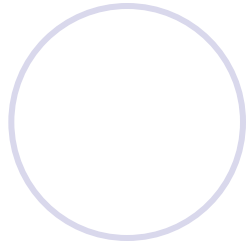
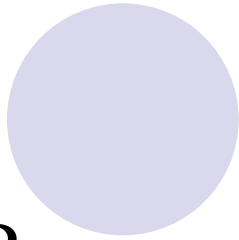
эконометристы

считают,

что

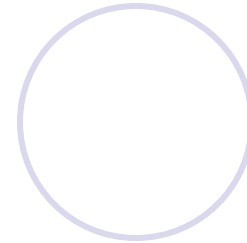
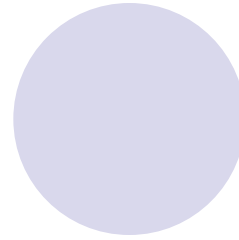
необходимо исключить «лишние» регрессоры,  
которые могут служить причиной  
мультиколлинеарности.

Но при этом могут возникнуть новые  
проблемы.



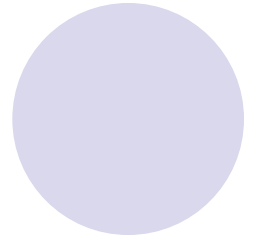
Во-первых,

не



всегда

ясно,



какие

переменные являются «лишними».

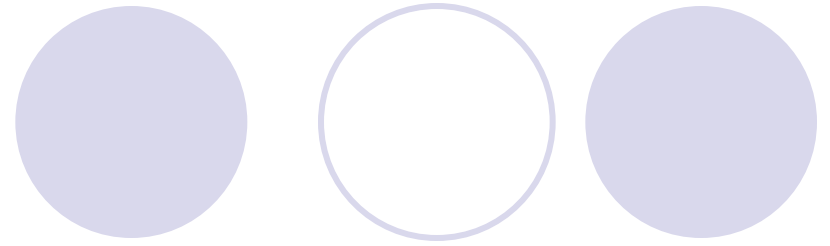
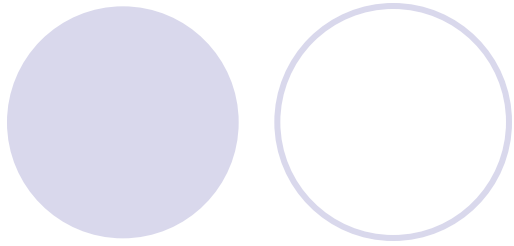
Во-вторых,

удаление

независимых

переменных может значительно отразиться на

содержательном смысле модели.



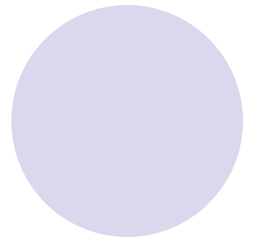
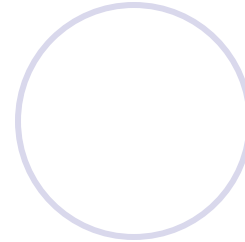
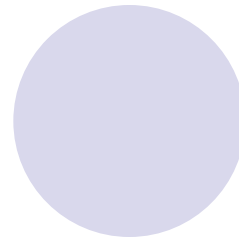
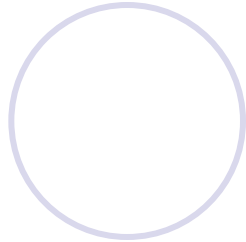
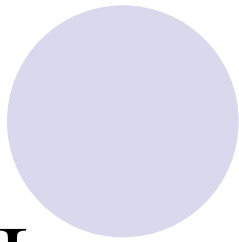
В-третьих, удаление переменных, которые реально влияют на изучаемую зависимую переменную, приводит к смещению МНК-оценок.



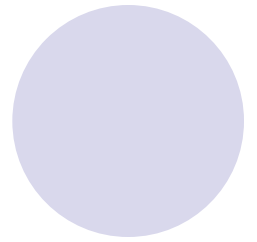
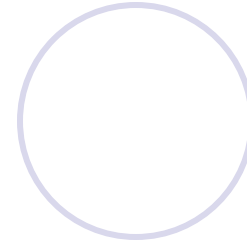
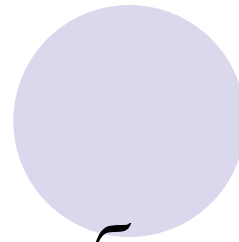
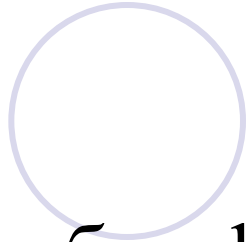
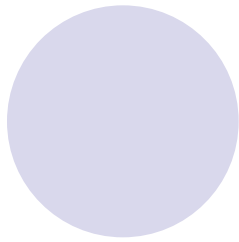
Теоретически регрессионная модель

позволяет учесть любое число факторов,  
практически в этом нет необходимости.

Отбор факторов проводится на основе  
качественного теоретико-экономического  
анализа.



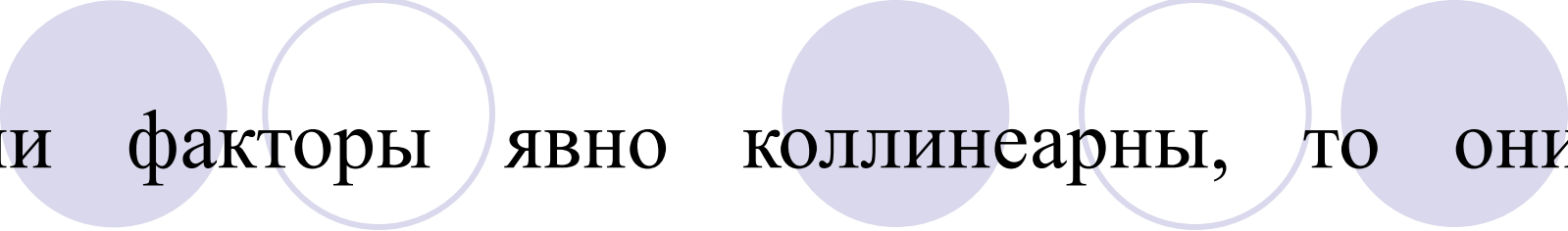
Но теоретический анализ не всегда  
позволяет однозначно ответить на  
вопрос о количественной взаимосвязи  
рассматриваемых признаков и  
целесообразности включения фактора  
в модель.



Поэтому отбор факторов обычно проводится  
в два этапа:

1. Отбираются факторы, исходя из сущности проблемы.
2. На основе матрицы парных коэффициентов корреляции и определения t-статистик для параметров регрессии.





Если факторы явно коллинеарны, то они дублируют друг друга и один из них рекомендуется исключить из регрессии.

Предпочтение при этом отдается не фактору, более тесно связанному с результатом, а тому фактору, который при достаточно тесной связи с результатом имеет наименьшую тесноту с другими факторами.



# Частные коэффициенты корреляции

Для решения проблемы коллинеарности можно использовать частные коэффициенты корреляции, которые характеризуют тесноту связи между результатом и регрессором при фиксированном влиянии других факторов.



$$r_{yx_1/x_2} = \frac{r_{yx_1} - r_{yx_2} * r_{x_1x_2}}{\sqrt{(1 - r_{yx_2}^2)(1 - r_{x_1x_2}^2)}}$$

Исключаем тот регрессор, для которого частный коэффициент наименьший, так как учтено взаимное влияние регрессоров.

# Коэффициент множественной корреляции, множественный коэффициент детерминации.

Коэффициент множественной корреляции используется для оценки тесноты связи между зависимой переменной и всеми регрессорами, включенными в модель.

$$R = \sqrt{\frac{\sigma^2_{\text{факт}}}{\sigma^2_{\text{общ}}}} = \sqrt{1 - \frac{\sigma^2_{\text{ост}}}{\sigma^2_{\text{общ}}}}$$

$$R \in [0;1]$$



$R^2$  – коэффициент множественной детерминации.

$R^2 * 100$  – доля вариации  $y$ , обусловленная включенными в модель факторами.

$(1 - R^2) * 100$  – доля вариации  $Y$ , обусловленная не включенными в модель факторами.

Проверка статистической значимости множественного коэффициента корреляции осуществляется также как и в парном анализе. Фактическое значение статистики Фишера определяется по формулам:

$$F_{\phi} = \frac{R^2}{1 - R^2} * \frac{n - k}{k - 1}$$

$n$  – размер выборки,  
 $k$  – общее число параметров, оцениваемых в уравнении.

$$F_{\phi} = \frac{R^2 / k}{(1 - R^2) \div (n - k - 1)}$$

$n$  – размер выборки,  
 $k$  – число независимых переменных.

# Стандартизированное уравнение множественной регрессии.

Существует другой подход к построению множественной регрессии – уравнение регрессии в стандартизированном масштабе. Для этого введем стандартизированные переменные

$$Z_y, Z_{x_1}, \dots, Z_{x_k}$$

$$Z_y = \frac{(y - \bar{y})}{\sigma_y}$$

$$Z_{x_k} = \frac{(x_k - \bar{x})}{\sigma_{x_k}}$$

Для этих переменных среднее значение равно 0, а среднее квадратическое отклонение равно 1.

$$Z_y = b_1 Z_{x_1} + b_2 Z_{x_2} + \dots + b_k Z_{x_k} + U$$

К этому уравнению можно применить МНК.  
Система:

$$r_{yx_1} = \beta_1 + \beta_2 r_{x_1 x_2} + \beta_3 r_{x_1 x_3} + \dots + \beta_k r_{x_1 x_k}$$

$$r_{yx_2} = \beta_1 r_{x_2 x_1} + \beta_2 + \beta_3 r_{x_2 x_3} + \dots + \beta_k r_{x_2 x_k}$$

...

$$r_{yx_k} = \beta_1 r_{x_k x_1} + \beta_2 r_{x_k x_2} + \beta_3 r_{x_k x_3} + \dots + \beta_k$$





$\beta$  – стандартизированные коэффициенты регрессии.

Данные коэффициенты сравнимы между собой и можно ранжировать факторы по силе воздействия на результат.

Стандартизированный коэффициент регрессии – показывает, на сколько средних квадратических отклонений изменится результат, если соответствующий фактор изменится на 1 сигма при неизменной величине остальных факторов.



Пример: Пусть функция издержек производства  $Y$  (тыс.руб.) характеризуется уравнением вида:

$$\hat{Y} = 200 + 1,2x_1 + 1,1x_2$$

где  $X_1$  – основные производственные фонды (тыс.руб.)

$X_2$  – численность занятых в производстве (чел.)



Построим уравнение в стандартизированном масштабе:

$$Z_y = 0,5 \cdot Z_{x_1} + 0,8 \cdot Z_{x_2}$$