



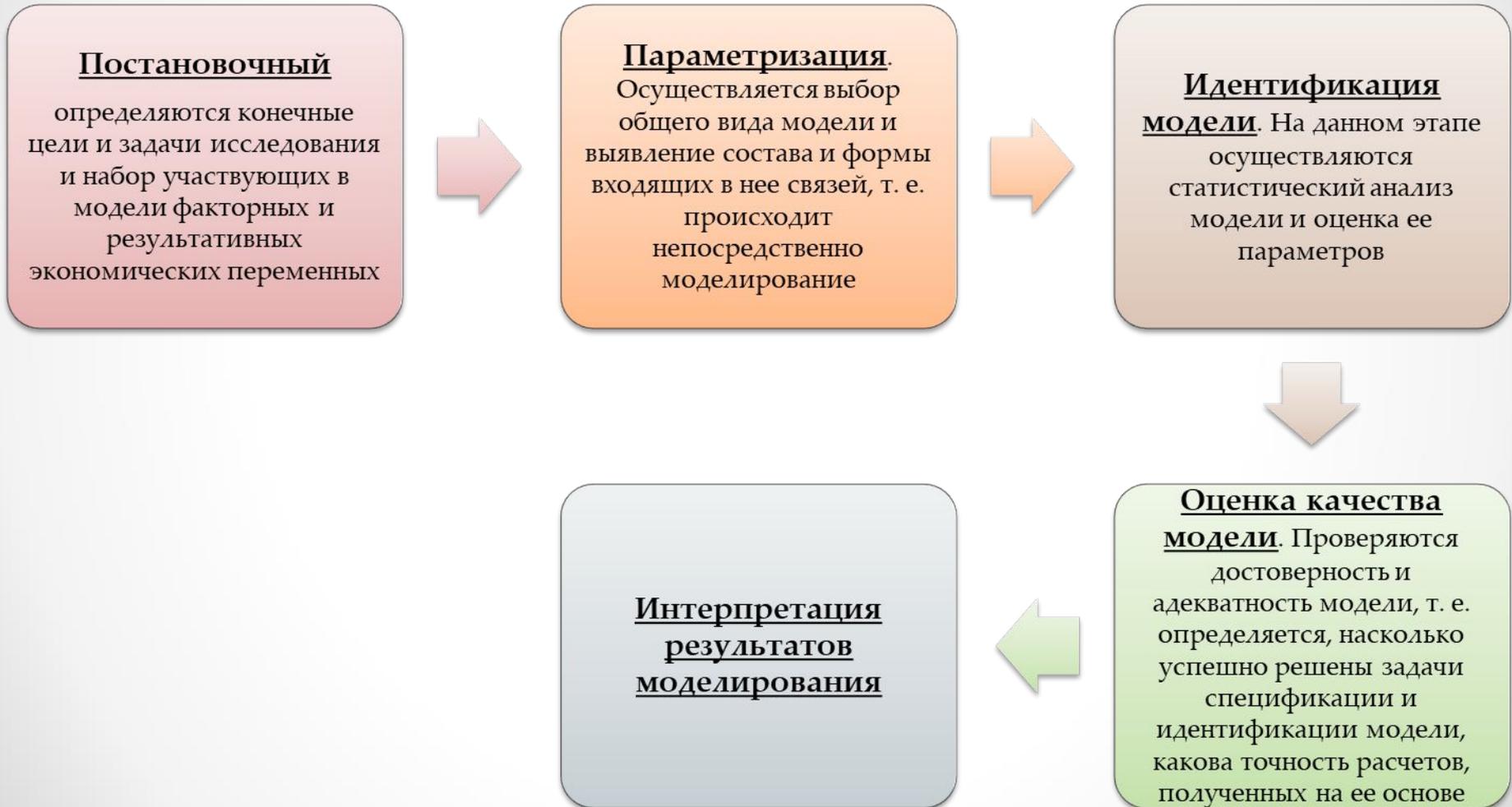
дисциплина:

**Современные методы
статистического анализа
кадастровых данных**

*к.э.н., профессор кафедры землеустройства и земельного
кадастра*

Яроцкая Елена Вадимовна

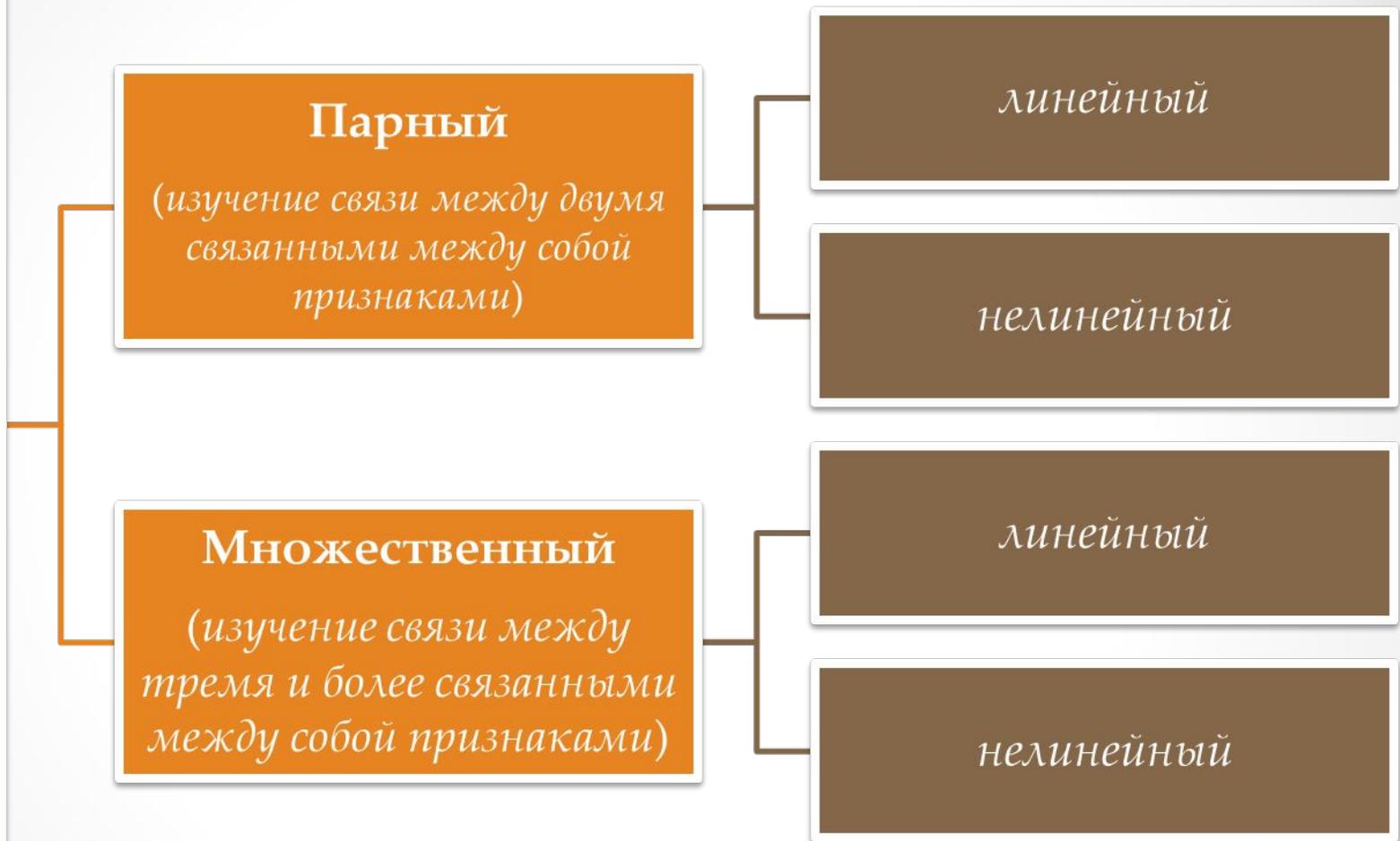
Этапы построения моделей



Регрессионный анализ заключается в определении аналитического выражения связи (в определении функции), в котором изменение одной величины (результативного признака) обусловлено влиянием независимой величины (факторного признака).

Количественно оценить данную взаимосвязь можно с помощью построения уравнения регрессии или регрессионной функции.

РЕГРЕССИОННЫЙ АНАЛИЗ



Задачи регрессионного анализа

Установление формы зависимости (линейная и нелинейная)

Определение функции регрессии в виде математического уравнения того или иного типа и установление влияния объясняющих переменных на зависимую переменную

Оценка неизвестных значений зависимой переменной. С помощью функции регрессии можно воспроизвести значения зависимой переменной внутри интервала заданных значений объясняющих переменных или оценить течение процесса вне заданного интервала

Линейная регрессия — используемая в статистике регрессионная модель зависимости одной (объясняемой, зависимой) переменной y от другой или нескольких других переменных (факторов, регрессоров, независимых переменных) x с линейной функцией зависимости.

Линейная парная модель наблюдений

$$y_i = (\alpha + \beta \cdot x_i) + \varepsilon_i, \quad i = 1, \dots, n.$$

если α и β — «истинные» значения параметров линейной модели СВЯЗИ, ТО

$$\varepsilon_i = y_i - (\alpha + \beta \cdot x_i)$$

ε представляет собой случайный член или *ошибку* в i -ом наблюдении

Случайная величина ε

характеризует отклонение
реального значения
результативного признака от
теоретического.

*Влияет на не учтённые в модели факторы,
случайных ошибок и особенностей измерений.*

Причины возникновения случайной ошибки:

- **Невключение объясняющих переменных.** Т.е. существуют другие факторы, влияющие на y , которые не учтены в уравнении. Влияние их приводит к тому, что точки не лежат на одной прямой.
- **Агрегирование переменных.** Рассматриваемая зависимость является попыткой объединить некоторое число объектов, которые, возможно, обладают различными характеристиками.
- **Неправильное описание структуры модели.** Т.е., если зависимость относится к данным о временном ряде, то значение y может зависеть не от фактического значения x , а от значения, которое ожидалось в предыдущем периоде. Если ожидаемое и фактическое значение тесно связаны, то будет казаться, что между y и x существует зависимость, но это будет лишь аппроксимация, и расхождение вновь будет связано с наличием случайного члена.
- **Неправильная функциональная спецификация.** Т.е. функциональное соотношение между y и x может быть определено неправильно.
- **Ошибки измерения.** Если в измерении одной или более взаимосвязанных переменных имеются ошибки, то наблюдаемые значения не соответствуют такому соотношению, и существующие расхождения будут увеличивать значения остаточного члена.

Параметр β - коэффициент регрессии - на сколько в среднем изменится результативный признак y при изменении факторного признака x на единицу своего измерения.

Знак параметра β в уравнении парной регрессии указывает на направление связи.

Если, $\beta > 0$, то связь между изучаемыми показателями прямая, т. е. с увеличением факторного признака x увеличивается и результативный признак, и наоборот.

Если $\beta < 0$, то связь между изучаемыми показателями обратная, т. е. с увеличением фактора x результат уменьшается, и наоборот.

Значение параметра α в уравнении парной регрессии трактуется как среднее значение результативного признака y при условии, что факторный признак x равен нулю. Такая трактовка параметра α возможна только в том случае, если значение $x = 0$ имеет смысл.

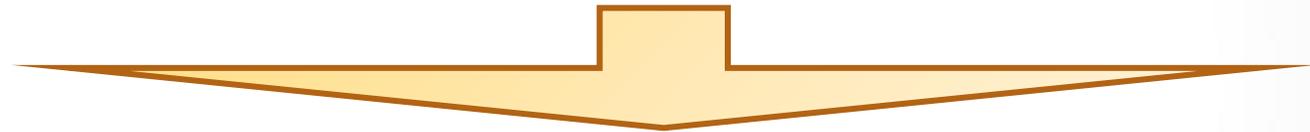
Метод наименьших квадратов (МНК) -

позволяет получить такие оценки параметров, при которых сумма квадратов отклонений фактических значений результативного признака y от теоретических \hat{y}_x минимальна

$$\sum (y - \hat{y}_x)^2 \rightarrow \min$$

Для линейных и нелинейных уравнений, приводимых к линейным, решается следующая система относительно a и b .

$$\begin{cases} na + b \sum x = \sum y, \\ a \sum x + b \sum x^2 = \sum yx. \end{cases}$$



$$b = \frac{\overline{xy} - \bar{y} \cdot \bar{x}}{\overline{x^2} - \bar{x}^2}$$

$$a = \bar{y} - b\bar{x}$$

Нелинейная регрессия — это вид регрессионного анализа, в котором экспериментальные данные моделируются функцией, являющейся нелинейной комбинацией параметров модели и зависящей от одной и более независимых переменных

Классы нелинейных регрессий

нелинейные относительно
включенных в анализ
объясняющих переменных,
но линейные по
оцениваемым параметрам

полиномы разных степеней

$$y = a + b_1 \cdot x + b_2 \cdot x^2 + b_3 \cdot x^3 + \varepsilon$$

равносторонняя гиперболоа

Метод наименьших квадратов (МНК) -
позволяет получить такие оценки параметров,
при которых сумма квадратов отклонений
фактических значений резульативного
признака y от теоретических \hat{y}_x минимальна

нелинейные по
оцениваемым параметрам

степенная

$$y = a \cdot x^b \cdot \varepsilon$$

показательная

$$y = a \cdot b^x \cdot \varepsilon$$

экспоненциальная

$$y = e^{a+b \cdot x} \cdot \varepsilon$$

Оценка значимости построенной модели парной регрессии

```
graph TD; A[Оценка значимости построенной модели парной регрессии] --> B[Адекватность модели (качество)]; A --> C[Значимость коэффициентов уравнения регрессии];
```

**Адекватность модели
(качество)**

**Значимость
коэффициентов
уравнения регрессии**

Адекватность модели (качество)

1) *коэффициент аппроксимации* – среднее отклонение расчетных значений от фактических:

$$\bar{A} = \frac{1}{n} \sum \left| \frac{y - \hat{y}}{\hat{y}} \right| \cdot 100\%$$

Допустимый предел значений – не более 8-10%.

Адекватность модели (качество)

2) доля дисперсии, объясняемую регрессией, в общей дисперсии результативного признака y характеризует индекс детерминации R^2 :

Метод наименьших квадратов (МНК) - позволяет получить такие оценки параметров, при которых сумма квадратов отклонений фактических значений результативного признака y от теоретических \hat{y}_x минимальна

чем ближе к 1, тем лучше качество модели

Адекватность модели (качество)

3) *F-тест* – оценивание качества уравнения регрессии – состоит в проверке гипотезы H_0 о статистической незначимости уравнения регрессии и показателя тесноты СВЯЗИ.

Для этого выполняется сравнение фактического $F_{\text{факт}}$ и критического (табличного) $F_{\text{табл}}$ значений F-критерия Фишера.

$$F_{\text{факт}} = \frac{\sum (\hat{y} - y)^2 / m}{\sum (y - \hat{y})^2 / (n - m - 1)} = \frac{r_{xy}^2}{1 - r_{xy}^2} (n - 2)$$

где n – число единиц совокупности,
 m – число параметров при переменных x .

$F_{табл}$ - это максимально возможное значение критерия под влиянием случайных факторов при данных степенях свободы и уровне значимости α

Уровень значимости α - вероятность отвергнуть правильную гипотезу при условии, что она верна. Обычно α принимается равной 0.05 или 0.01

$F_{табл}(\alpha; k_1; k_2)$ определяется по таблице и зависит от уровня значимости, числа степеней свободы $k_1 = m$ и числа степеней свободы $k_2 = n - m - 1$

Если $F_{табл} < F_{факт}$ то гипотеза H_0 - гипотеза о случайной природе оцениваемых характеристик отклоняется и признается их статистическая значимость и надежность. В противном случае признается их статистическая не значимость и не надежность

Значимость коэффициентов уравнения регрессии

1) *t*-критерий Стьюдента. Выдвигается гипотеза о случайной природе показателей, т.е. о незначимом их отличии от нуля.

t-критерия Стьюдента рассчитываются для параметров линейной регрессии и коэффициента корреляции

$$t_a = \frac{a}{m_a}$$

$$t_b = \frac{b}{m_b}$$

$$t_r = \frac{r}{m_r}$$

Случайные ошибки параметров линейной регрессии и коэффициента корреляции

$$m_b = \sqrt{\frac{\sum (y - \hat{y})^2 / (n - 2)}{\sum (x - \bar{x})^2}} = \sqrt{\frac{S_{\text{ост}}^2}{\sum (x - \bar{x})^2}} = \frac{S_{\text{ост}}}{\sigma_x \sqrt{n}}$$

$$m_{r_{xy}} = \sqrt{\frac{1 - r_{xy}^2}{n - 2}}$$

$$m_a = \sqrt{\frac{\sum (y - \hat{y})^2 \cdot \sum x^2}{(n - 2) \cdot n \cdot \sum (x - \bar{x})^2}} = \sqrt{\frac{S_{\text{ост}}^2 \cdot \sum x^2}{n^2 \sigma_x^2}} = \frac{S_{\text{ост}}}{\sigma_x n} \sqrt{\sum x^2}$$

Оценка значимости коэффициентов регрессии и корреляции с помощью t -критерия Стьюдента проводится путем сопоставления их значений с величиной случайной ошибки

Сравнивая фактическое и критическое (табличное) значения t -статистики $t_{\text{табл}}$ и $t_{\text{факт}}$ - принимаем или отвергаем гипотезу H_0

Если $t_{\text{табл}} < t_{\text{факт}}$ то H_0 отклоняется, т.е. a, b, r_{xy} не случайно отличаются от нуля и сформировались под влиянием систематически действующего фактора x

Если $t_{\text{табл}} > t_{\text{факт}}$ то гипотеза H_0 не отклоняется и признается случайная природа формирования a, b, r_{xy} .

Значимость коэффициентов уравнения регрессии

2) *Доверительный интервал* – предельные значения статистической величины, которая с заданной доверительной вероятностью будет находиться в этом интервале при выборке большего объема.

определяем предельную ошибку Δ для каждого показателя

$$\Delta_a = t_{\text{табл}} \cdot m_a$$

$$\Delta_b = t_{\text{табл}} \cdot m_b$$

Значимость коэффициентов уравнения регрессии

расчет доверительных интервалов

$$\gamma_a = a \pm \Delta_a$$

$$\gamma_{a_{\min}} = a - \Delta_a$$

$$\gamma_{a_{\max}} = a + \Delta_a$$

$$\gamma_b = b \pm \Delta_b$$

$$\gamma_{b_{\min}} = b - \Delta_b$$

$$\gamma_{b_{\max}} = b + \Delta_b$$

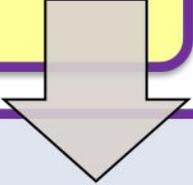
Значимость коэффициентов уравнения регрессии

Если в границы доверительного интервала попадает ноль, т. е. нижняя граница отрицательная, а верхняя положительна, то оцениваемый параметр принимается нулевым, так как он не может одновременно принимать и положительное, и отрицательное значения.

Множественная регрессия - изучение связи между тремя и более связанными между собой признаками. Требуется определить аналитическое выражение связи между признаком y и объясняющими переменными x_1, x_2, \dots, x_n в виде $y = f(x_1, x_2, \dots, x_n)$.

Этапы построения модели множественной регрессии

Выбор формы связи



Отбор факторных признаков



Обеспечение достаточного объема совокупности для получения несмещенных оценок

Линейная множественная модель наблюдений

$$y = \alpha_0 + \alpha_1 X_1 + \alpha_2 X_2 + \dots + \alpha_n X_n$$

α - значения параметров линейной модели связи

Оценка значимости построенной модели

1) Коэффициент множественной детерминации

$$\tilde{R}^2 = 1 - (1 - R^2) \frac{n-1}{n-m-1},$$

$R^2_{y x_1 x_2 \dots x_p}$

квадрат индекса множественной корреляции

где n – число наблюдений,
 m – число факторов

4) Мультиколлинеарность –

понятие, которое используется для описания проблемы, когда нестрогая линейная зависимость между факторами приводит к получению ненадежных оценок регрессии

4) F-критерий Фишера

$$F = \frac{R^2}{1 - R^2} \cdot \frac{n - m - 1}{m}$$

$F_{табл}$ - это максимально возможное значение критерия под влиянием случайных факторов при данных степенях свободы и уровне значимости α

Уровень значимости α - вероятность отвергнуть правильную гипотезу при условии, что она верна.
Обычно α принимается равной 0.05 или 0.01

$F_{табл}(\alpha; k_1; k_2)$ определяется по таблице и зависит от уровня значимости, числа степеней свободы $k_1 = m$ и числа степеней свободы $k_2 = n - m - 1$

Если $F_{табл} < F_{факт}$ то гипотеза H_0 - гипотеза о случайной природе оцениваемых характеристик отклоняется и признается их статистическая значимость и надежность. В противном случае признается их статистическая не значимость и не надежность

Значимость коэффициентов уравнения регрессии

1) *t*-критерий Стьюдента. Выдвигается гипотеза о случайной природе показателей, т.е. о незначимом их отличии от нуля.

$$t_{b_i} = \frac{b_i}{m_{b_i}} = \sqrt{F_{x_i}}$$

где
 m_{b_i} - средняя квадратическая ошибка коэффициента регрессии b_i ,

$$m_{b_i} = \frac{\sigma_y \cdot \sqrt{1 - R_{yx_1 \dots x_p}^2}}{\sigma_{x_i} \cdot \sqrt{1 - R_{x_i x_1 \dots x_p}^2}} \cdot \frac{1}{\sqrt{n - m - 1}}$$

Нелинейные множественные регрессионные модели

Степенная:

$$y = \alpha_0 x_1^{\alpha_1} x_2^{\alpha_2} \dots x_n^{\alpha_n}$$

Показательная

$$y = e^{\alpha_0 + \alpha_1 x_1 + \dots + \alpha_n x_n}$$

Параболическая:

$$y = \alpha_0 + \alpha_1 x_1^2 + \dots + \alpha_n x_n^2$$

Гиперболическая:

$$y = \alpha_0 + \frac{\alpha_1}{x_1} + \dots + \frac{\alpha_n}{x_n}$$



Кубанский государственный
аграрный университет

Землеустроительный
факультет

Благодарю за внимание!