# Обработка результатов моделирования методами корреляционно-регрессионного анализа

НАЦИОНАЛЬНЫЙ ИССЛЕДОВАТЕЛЬСКИЙ МОСКОВСКИЙ ГОСУДАРСТВЕННЫЙ
СТРОИТЕЛЬНЫЙ УНИВЕРСИТЕТ

# Основные типы кривых, используемые при количественной оценке связей между двумя переменными

$$а)\ \hat{y}_x = a + b \cdot x;$$

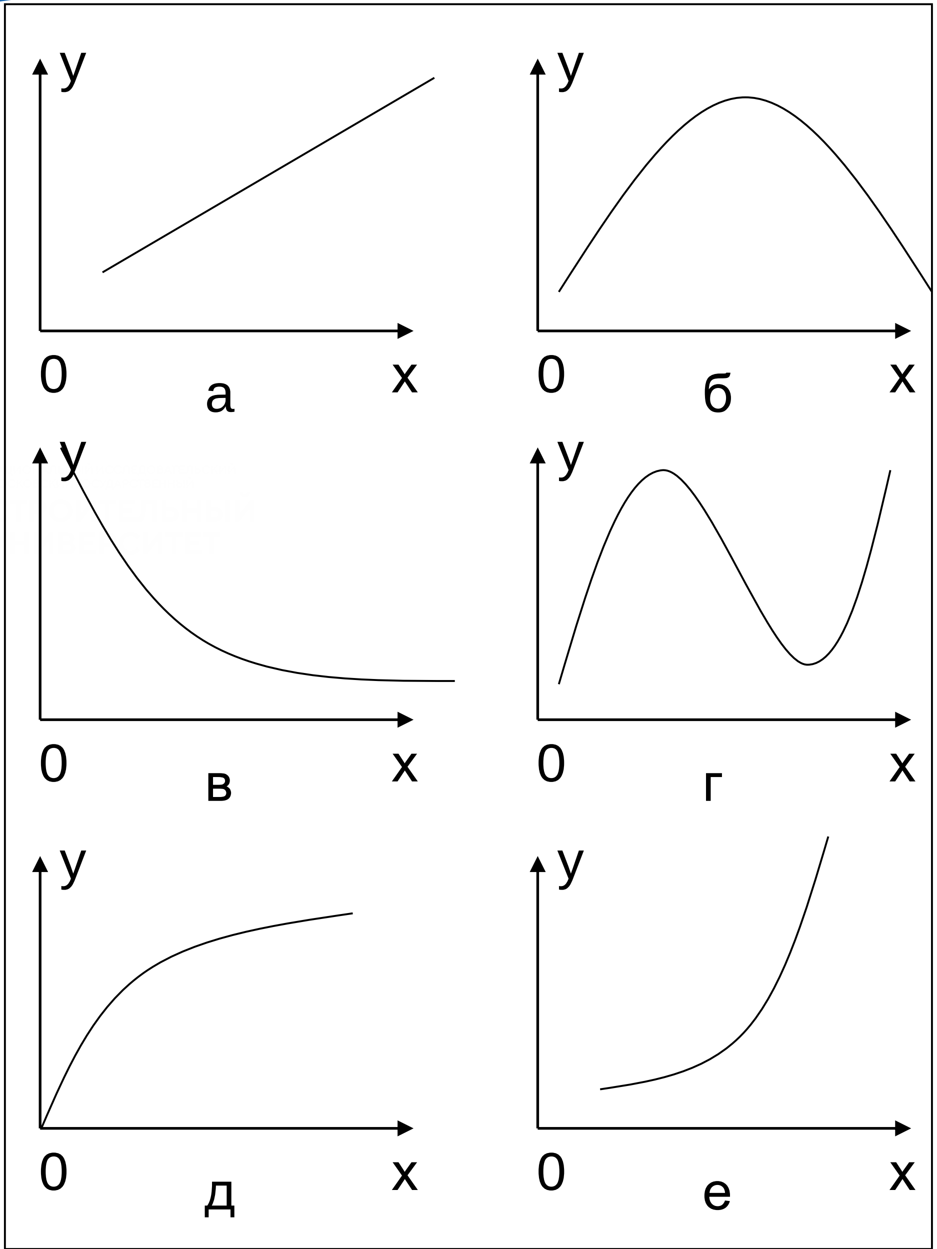
$$б)\ \hat{y}_x = a + b \cdot x + c \cdot x^2$$

$$в)\ \hat{y}_x = a + b / x;$$

$$г)\ \hat{y}_x = a + b \cdot x + c \cdot x^2 + d \cdot x^3$$

$$д)\ \hat{y}_x = a \cdot x^b;$$

$$е)\ \hat{y}_x = a \cdot b^x$$



# Парная линейная регрессия

Парная регрессия – уравнение связи двух переменных *x* и *y*

$$y = \widehat{f}(x)$$

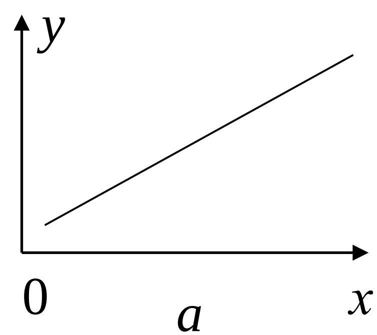
где *x* - независимая, объясняющая переменная (признак-фактор),

*y* - зависимая переменная (результативный признак).

- Парная регрессия
  - Линейная
  - Нелинейная

**Линейная регрессия**

$$\widehat{y} = a + bx$$



**Нелинейные по независимым переменным**

- Полиномы
- Гипербола
- Полулогарифмическая функция

**Нелинейные по оцениваемым параметрам**

- Степенная
- Показательная
- Экспоненциальная

# Формулы для нахождения параметров линейной регрессии

$$\widehat{y} = a + bx$$

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

$$a = \overline{y} - b \cdot \overline{x}$$

$$\overline{x} = \frac{1}{n}\sum x \qquad \overline{y} = \frac{1}{n}\sum y \qquad \overline{y \cdot x} = \frac{1}{n}\sum y \cdot x \qquad \overline{x^2} = \frac{1}{n}\sum x^2$$

# Интерпретация уравнения регрессии

- *b* – коэффициент регрессии
  - Показывает среднее изменение результата с изменением фактора на одну единицу
- *a* – может не иметь экономического смысла
  - Формально *a* – значение результата при нулевом значении фактора

*Пример.* $y=3000+2x$

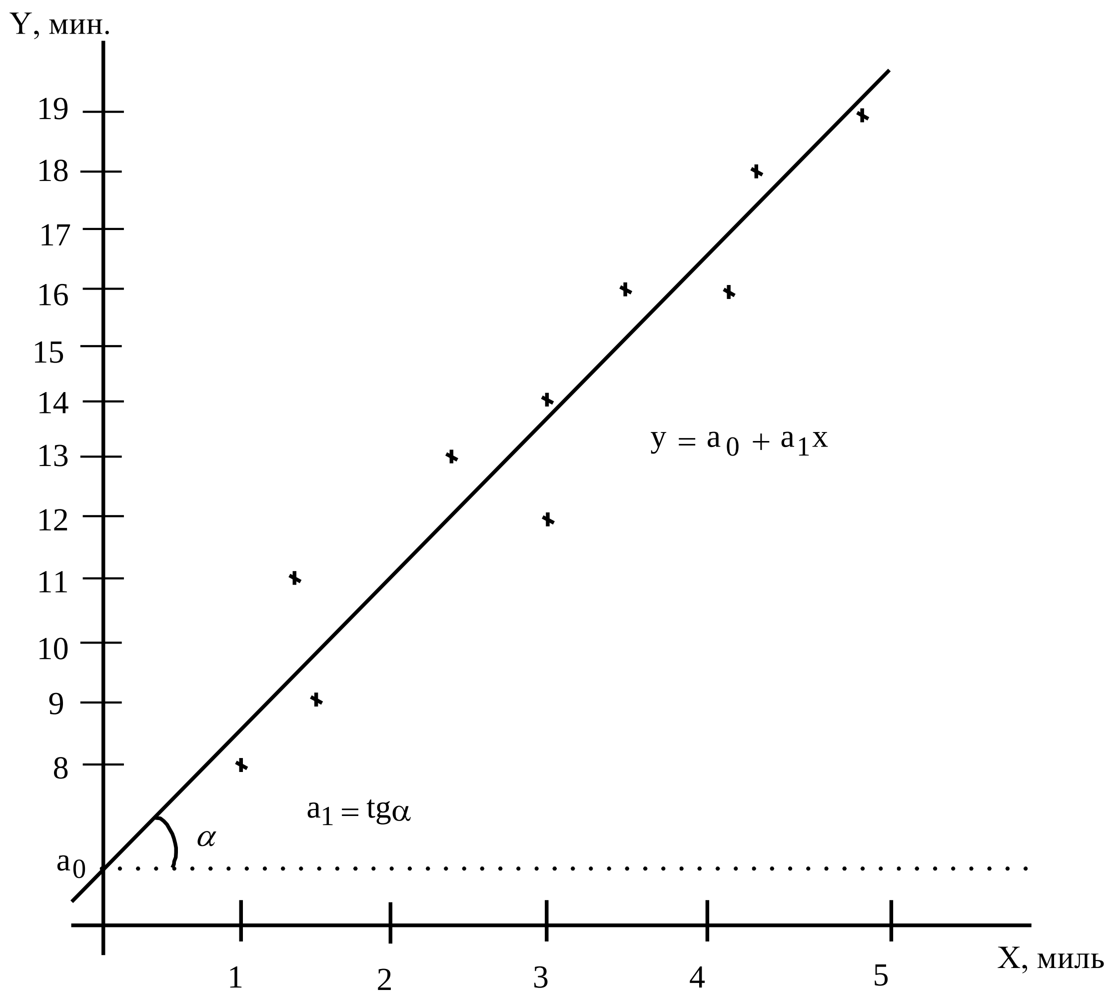
*y – издержки (тыс. руб.) x – количество единиц продукции*

Некоторая фирма занимается поставками различных грузов на короткие расстояния внутри города. Перед менеджером стоит задача оценить стоимость таких услуг, зависящую от затрачиваемого на поставку времени. В качестве наиболее важного фактора, влияющего на время поставки, менеджер выбрал пройденное расстояние. Были собраны исходные данные о десяти поставках

| | | | | | | | | | | |
|---|---|---|---|---|---|---|---|---|---|---|
| Расстояние, миль | 3,5 | 2,4 | 4,9 | 4,2 | 3,0 | 1,3 | 1,0 | 3,0 | 1,5 | 4,1 |
| Время, мин | 16 | 13 | 19 | 18 | 12 | 11 | 8 | 14 | 9 | 16 |

| | | | | | | | | | | |
|---|---|---|---|---|---|---|---|---|---|---|
| Расстояние, миль | 3,5 | 2,4 | 4,9 | 4,2 | 3,0 | 1,3 | 1,0 | 3,0 | 1,5 | 4,1 |
| Время, мин | 16 | 13 | 19 | 18 | 12 | 11 | 8 | 14 | 9 | 16 |

# Поле корреляции



$\sum y_i =$

Расстояние, миль - $x$

Время, мин - $y$

Вспомогательная таблица

| $x_i$ | $y_i$ | $x_i^2$ | $x_i y_i$ | $y_i^2$ |
|---|---|---|---|---|
| 3,5 | 16 | 12,25 | 56,00 | 256 |
| 2,4 | 13 | 5,76 | 31,2 | 169 |
| 4,9 | 19 | 24,01 | 93,1 | 361 |
| 4,2 | 18 | 17,64 | 75,60 | 324 |
| 3,0 | 12 | 9,00 | 36,00 | 144 |
| 1,3 | 11 | 1,69 | 14,30 | 121 |
| 1,0 | 8 | 1,00 | 8,00 | 64 |
| 3,0 | 14 | 9,00 | 42,00 | 196 |
| 1,5 | 9 | 2,25 | 13,50 | 81 |
| 4,1 | 16 | 16,81 | 65,60 | 256 |
| $\sum = 28{,}9$ | $\sum = 136$ | $\sum = 99{,}41$ | $\sum = 435{,}30$ | $\sum = 1972$ |

$$b = \frac{\frac{1}{10} \cdot 435{,}30 - \frac{1}{10} 136 \cdot \frac{1}{10} 28{,}9}{\frac{1}{10} \cdot 99{,}41 - (\frac{1}{10} 28{,}9)^2} = 2{,}66; \;\; a = \frac{1}{10} \cdot \left(136 - 2{,}660 \cdot 28{,}9\right) = 5{,}913.$$

$$y = 5{,}913 + 2{,}66\,x.$$

При увеличении расстояния на 1 милю время увеличится на 2,66 мин

# Наличие связи между переменными

- Показатель тесноты связи $r_{xy}$ – линейный коэффициент корреляции

$$r_{xy} = \frac{\overline{yx} - \overline{y} \cdot \overline{x}}{\sqrt{(\overline{x^2} - \overline{x}^2)(\overline{y^2} - \overline{y}^2)}}$$

$$-1 \le r_{xy} \le 1$$

- Коэффициент детерминации $R^2 = r_{xy}^2$

# Шкала Чеддока

| Теснота связи | Значение коэффициента корреляции при наличии: | |
|---|---|---|
| | Прямой связи | Обратной связи |
| Слабая | 0,1–0,3 | (–0,3)–(–0,1) |
| Умеренная | 0,3–0,5 | (–0,5)–(–0,3) |
| Заметная | 0,5–0,7 | (–0,7)–(–0,5) |
| Высокая | 0,7–0,9 | (–0,9)–(–0,7) |
| Весьма высокая | 0,9–1 | (–1)–(–0,9) |

$r_{xy}=1,96$
$r_{xy}=0,03$
$r_{xy}=0,49$
$r_{xy}= -0,81$
$r_{xy}= -0,99$

1. допущена ошибка в вычислениях
2. отсутствует какая-либо связь
3. слабая обратная связь
4. умеренная прямая связь
5. заметная прямая связь
6. высокая обратная связь
7. весьма высокая обратная связь

# Пример (продолжение)

$$r_{xy} = \frac{\frac{1}{10} \cdot 435{,}30 - \frac{1}{10}136 \cdot \frac{1}{10}28{,}9}{\sqrt{\frac{1}{10} \cdot 99{,}41 - (\frac{1}{10}28{,}9)^2}\sqrt{\frac{1}{10} \cdot 1972 - (\frac{1}{10}136)^2}} = 0{,}96.$$

$$R^2 = r_{xy}^2 = 0{,}96^2 = 0{,}92$$

Близость коэффициента корреляции к 1 указывает на тесную линейную связь между признаками.

Коэффициент детерминации показывает, что уравнением регрессии объясняется 92% дисперсии результативного признака, а на долю прочих факторов приходится лишь 8%.

# Прогнозирование по модели

***Прогнозное значение*** результата $y_p$ (точечный прогноз) определяется путем подстановки в уравнение регрессии

$$\hat{y} = a + bx$$

соответствующего (прогнозного) значения фактора $x_p$

Зависимость времени от расстояния имеет вид

$$y = 5{,}913 + 2{,}66\,x.$$

$$y_p(2) = 5{,}913 + 2{,}66 \cdot 2 = 11{,}2$$

**1**. Значение коэффициента корреляции равно 0,7. Чему равно значение коэффициента детерминации?

**2.** Для уравнения зависимости выручки от величины оборотных средств получено значение коэффициента детерминации, равное 0,6. Сколько процентов дисперсии обусловлено случайными факторами?

**3.** Для уравнения зависимости выручки от величины оборотных средств получено, что 65% дисперсии обусловлено случайными факторами Каково значение коэффициента детерминации?

# Нелинейные регрессии

Нелинейная регрессия

- Нелинейные по независимым переменным
- Независимые по оцениваемым параметрам

# Нелинейные по оцениваемым параметрам

## Нелинейные модели внутренне линейные

- Степенная
- Показательная
- Экспоненциальная
- Обратная
- Логистическая

$$\hat{y}_x = a \cdot x^b$$

$$\hat{y}_x = a \cdot b^x$$

$$\hat{y}_x = a\,\mathrm{e}^{b \cdot x}$$

$$\hat{y}_x = \frac{1}{a + b \cdot x}$$

$$\hat{y}_x = \frac{a}{1 + b \cdot \mathrm{e}^{-c \cdot x}}$$

## Нелинейные модели внутренне нелинейные

$$\hat{y}_x = a + b \cdot x^c \qquad \hat{y}_x = a \cdot \left(1 - \frac{1}{1 - x^b}\right)$$

# Линеаризация

- Парабола $$y_x = a + b \cdot x + c \cdot x^2$$

$$x = x_1,\ \ x^2 = x_2$$

- Гипербола $$y_x = a + b/x$$

$$X = {}^{1}\!/_{x}$$

$$b = \frac{\overline{yX} - \bar{y} \cdot \bar{X}}{\overline{X^2} - (\bar{X})^2}$$

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

- Полулогарифмическая функция $$y_x = a + b \cdot \ln x$$

$X$=ln$x$

$$b = \frac{\overline{yX} - \bar{y} \cdot \bar{X}}{\overline{X^2} - (\bar{X})^2}$$

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

# Нелинейные модели внутренне линейные

- Степенная $y=ax^b$

$Y=\ln y,\ X=\ln x,\ A=\ln a$

$$b=\frac{\overline{YX}-\overline{Y}\cdot\overline{X}}{\overline{X^2}-(\overline{X})^2}\qquad A=\overline{Y}-b\overline{X}$$

- Показательная $y=ab^x$

$Y=\ln y,\ B=\ln b,\ A=\ln a$

$$B=\frac{\overline{Yx}-\overline{Y}\cdot\overline{x}}{\overline{x^2}-(\overline{x})^2}\qquad A=\overline{Y}-B\overline{x}$$

- Экспоненциальная $y=ae^{bx}$

$Y=\ln y,\ A=\ln a$

$$b=\frac{\overline{Yx}-\overline{Y}\cdot\overline{x}}{\overline{x^2}-(\overline{x})^2}\qquad A=\overline{Y}-b\overline{x}$$

- Обратная $$y_x=\frac{1}{a+b\cdot x}$$

$Y=1/y$

$$b=\frac{\overline{Yx}-\overline{Y}\cdot\overline{x}}{\overline{x^2}-(\overline{x})^2}\qquad \grave{a}=\overline{Y}-b\overline{x}$$

# Показатель тесноты связи

- Индекс корреляции

$$\rho_{xy} = \sqrt{1 - \frac{\sum \left(y - y_x\right)^2}{\sum \left(y - \bar{y}\right)^2}}$$

Интервал изменения

$$0 \leq \rho_{xy} \leq 1$$

- Коэффициент детерминации $\rho_{xy}^2$