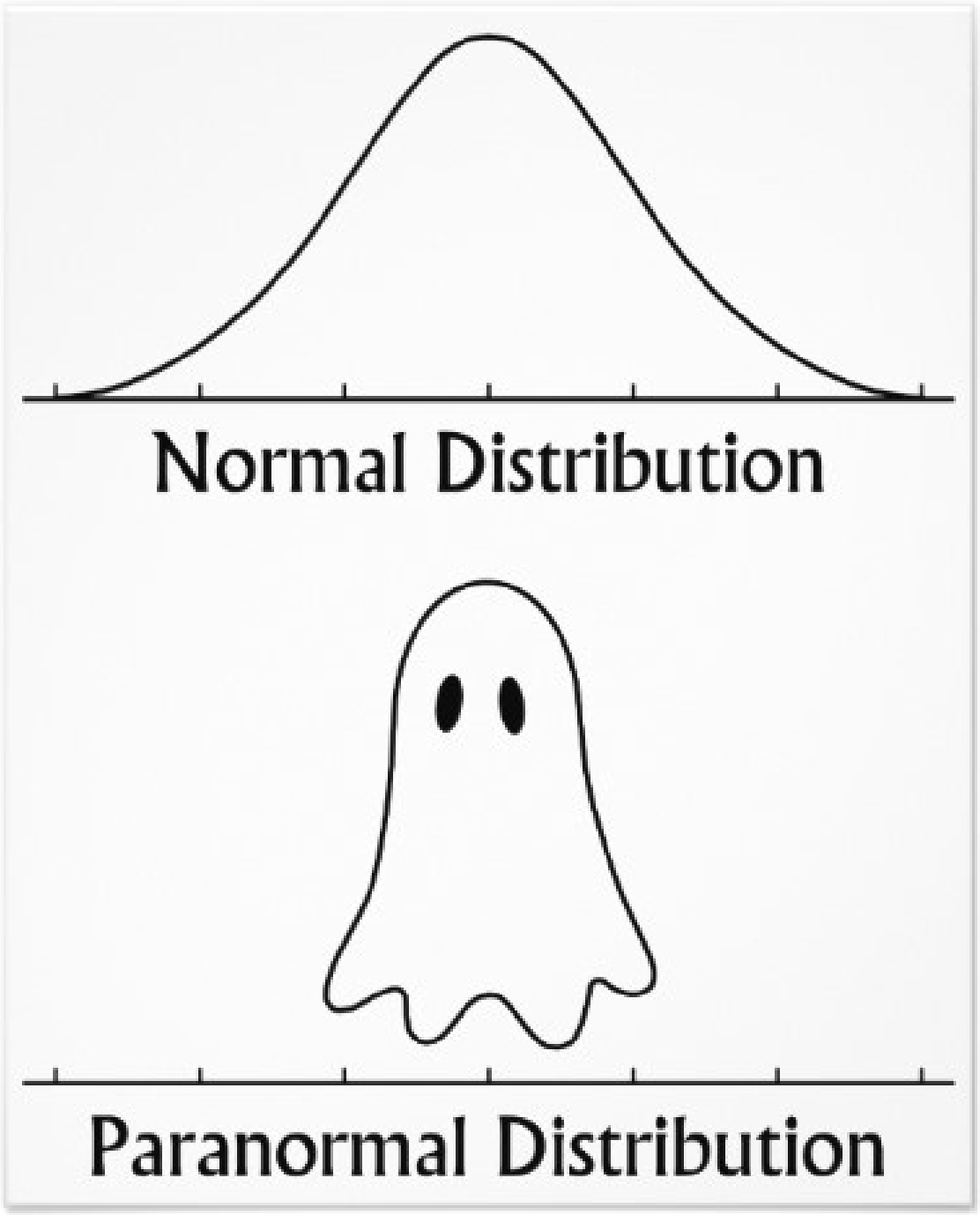

Normal Distribution
Paranormal Distribution

Регрессия-шпаргалка

Кафедра Автоматизации технологических процессов

Доц. Южанин В.В.

Об использовании регрессионной модели для описания реальных процессов

- Ошибка (шум) моделирует неучтенные факторы. Невозможность провести регрессионную линию через все точки. Нет 100% воспроизводимости. При одних и тех же значениях x величина y будет разная.
- Для объяснения такого рода данных вводят регрессионную модель, содержащую и детерминированный и случайный компонент в наблюдаемом параметре:

$$y(x) = x^T\beta + \varepsilon$$

- Переменные (они же факторы или регрессоры) x считаются известными точно, т.е. являются детерминированными величинами (в отличие от случайных откликов y).
- Наблюдаемые значения y содержит как детерминированную, так и случайную компоненту.
- В классической регрессии принимается, что шум нормальный с нулевым МО, одинаковой дисперсией для всех y(x). В жизни шум может не быть нормальным и может иметь разную дисперсию.

О задаче МНК и линейности ее решения относительно откликов Y

- Критерий МНК минимизирует сумму квадратов невязок между прогнозом и фактом за счет подбора коэффициентов регрессионной модели.

$$Q(\beta) = e^T e = (Y - X\beta)^T (Y - X\beta)$$

- Решение задачи метода наименьших квадратов – вектор коэффициентов регрессионной модели.

$$\hat{\beta} = arg \underbrace{min}_{\beta} Q(\beta)$$

- При решении задачи МНК из всех выборочных значений y составляется вектор Y = {y1, y2, … yn}’.
- Вектору фактических значений Y соответствует вектор прогноза, который рассчитывается через X-матрицу и вектор регрессионных коэффициентов.
- Решение задачи МНК:

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

- Несмотря на то, что минимизируется ***квадрат*** невязок, вектор коэффициентов регрессии при решении задачи МНК **линейно** зависит от вектора Y.
- Матрица $(X^T X)^{-1} X^T$ может быть вычислена, поскольку определяется значениями факторов x, которые в регрессии считаются детерминированными и известными.

О случайности оценок коэффициентов регрессии, восстановленных по фактическим данным и как следствие случайности прогноза.

- Оценка векторов коэффициентов регрессии отличается от выборки к выборке из-за случайности шума. Следовательно, вектор коэффициентов регрессии – случайный вектор, компоненты которого – случайные величины.
- Прогноз $\hat{y}(x) = x^T \hat{\beta}$ также случайная величина, т.к. зависит от $\hat{\beta}$.
- Если знать распределение $\hat{y}(x)$, то можно построить доверительный интервал прогноза, т.е. в итоге у нас будет не просто модель, но и численная характеристика ее погрешности.
- Очевидно, что для этого придется узнать и распределение $\hat{\beta}$.

О распределении оценок коэффициентов регрессии $\hat{\beta}$ и прогноза $\hat{y}(x)$ - 1

- В задаче МНК компоненты вектора Y подчиняются нормальному распределению, т.к. содержат нормальный шум. Все его линейные преобразования вида AY также будут нормальным случайными векторами.
- В решении задачи МНК матрица A = $(X'X)^{-1}X'$. Таким образом, $\hat{\beta} = A \cdot Y$ – нормальный случайный вектор.
- Все компоненты вектора $\hat{\beta}$ зависят от одного и того же вектора Y, поэтому между компонентами этого вектора появляется статистическая связь (корреляция).
- **В дальнейшем в прогнозе $\hat{y}(x) = x^T\hat{\beta}$ участвуют все компоненты вектора $\hat{\beta}$. Связь между компонентами $\hat{\beta}$ повлияет на распределение прогноза.**
- Обобщением понятия дисперсии для векторных случайных величин служит ковариационная матрица, которая содержит информацию как о собственно дисперсии компонент случайного вектора (диагональные элементы), так и о попарных коэффициентах ковариации между компонентами (недиагональные элементы).

О распределении оценок коэффициентов регрессии $\hat{\beta}$ и прогноза $\hat{y}(x)$ - 2

- Ковариационная матрица случайного вектора $\hat{\beta}$.

$$K = cov(\hat{\beta}) = E((\hat{\beta} - \beta)(\hat{\beta} - \beta)^T) = \sigma^2 (X^T X)^{-1}$$

- Дисперсия прогноза.

$$cov(\hat{y}(x)) = D(\hat{y}) = E\left(\underbrace{\hat{y}(x) - y_{\text{и}}(x)}_{e_{\text{и}}}\right)(\hat{y}(x) - y_{\text{и}}(x))^T$$

$$= x^T cov(\hat{\beta}) x = x^T K x$$

- В итоге (учитывая **несмещенность** оценок $\hat{\beta}$ и прогноза)

$$\hat{\beta}_i \sim N(\beta_i, K_{i,i})$$

$$\hat{y}(x) \sim N(x^T \beta, D(\hat{y}))$$

О доверительных интервалах коэффициентов регрессии $\hat{\beta}$ и прогноза $\hat{y}(x)$ при известной дисперсии шума.

- Итак, вроде бы готовы решить задачу о 95%-доверительном интервале прогноза $\hat{y}(x)$:

$$\hat{y}(x) - x^T\beta \sim N\left(0, D(\hat{y})\right)$$

- Тогда 95% значений данной СВ лягут в диапазоне

$$norminv\left(\frac{\alpha}{2}, 0, D(\hat{y})\right) \le \hat{y}(x) - x^T\beta \le norminv\left(1 - \frac{\alpha}{2}, 0, D(\hat{y})\right)$$

где $\alpha = 1 - 0.95 = 0.05$

- Далее учтем, что вследствие симметрии

$$norminv\left(1 - 1/\alpha, 0, D(\hat{y})\right) = -norminv\left(\alpha/2, 0, D(\hat{y})\right)$$

$$-norminv\left(1 - 1/\alpha, 0, D(\hat{y})\right) \le \hat{y}(x) - x^T\beta \le norminv\left(1 - 1/\alpha, 0, D(\hat{y})\right)$$

$$\hat{y}(x) - norminv\left(1 - 1/\alpha, 0, D(\hat{y})\right) \le x^T\beta \le \hat{y}(x) + norminv\left(1 - 1/\alpha, 0, D(\hat{y})\right)$$

О доверительных интервалах прогноза $\hat{y}(x)$ при неизвестной дисперсии шума

- Проблема в том, что дисперсия $D(\hat{y})$ зависит от дисперсии шума σ^2, которая нам неизвестна.
- Если построить оценку дисперсии шума s, то можно построить оценку дисперсии прогноза $\widehat{D}(\hat{y})$. Поскольку прогноз $\hat{y}(x)$ подчиняется нормальному распределению, то следующая величина подчиняется распределению Стьюдента:

$$\frac{\hat{y}(x) - x^T\beta}{\sqrt{\widehat{D}(\hat{y})}} \sim t_{n-k}$$

где k – порядок модели, $k = |\hat{\beta}|$ (немного забежали вперед)

- Построим 95%-доверительный интервал данной СВ:

$$tinv\left(\frac{\alpha}{2}, n-k\right) \le \frac{\hat{y}(x) - x^T\beta}{\sqrt{\widehat{D}(\hat{y})}} \le tinv\left(1-\frac{\alpha}{2}, n-k\right)$$

где $\alpha = 1 - 0.95 = 0.05$

- Далее учтем, что вследствие симметрии

$$tinv\left(1-\frac{\alpha}{2}, n-k\right) = -tinv\left(\frac{\alpha}{2}, n-k\right)$$

$$\hat{y}(x) - tinv\left(1-\frac{\alpha}{2}, n-k\right) D(\hat{y}) \le x^T\beta \le \hat{y}(x) + tinv\left(1-\frac{\alpha}{2}, n-k\right) D(\hat{y})$$

Об оценке дисперсии прогноза, оценке ковариационной матрицы коэффициентов регрессии и несмещенной оценке дисперсии шума

- Дисперсия прогноза $D(\hat{y})$ зависит от ков. матрицы $K = cov(\hat{\beta})$, которая уже зависит от дисперсии шума.
- Пусть мы решили (непростой!) вопрос с тем как рассчитать оценку дисперсии шума s^2. Можно показать, что несмещенная оценка дисперсии шума имеет вид:

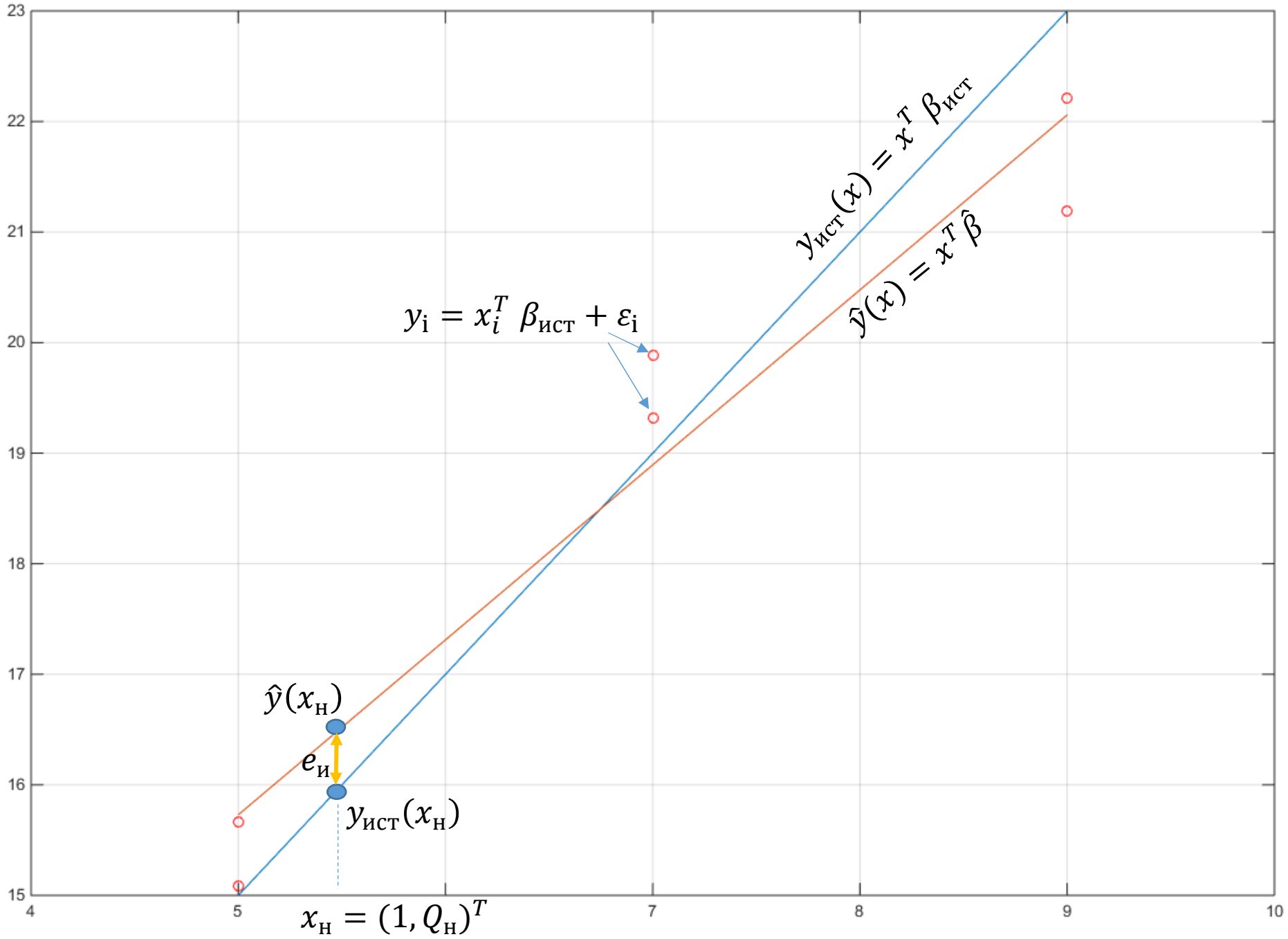
$$s^2 = \frac{Q(\hat{\beta})}{n-k}$$

где *n-k* – количество степеней свободы в распределении хи-квадрат.

- Далее заменим в известных формулах для $D(\hat{y})$ и $cov(\hat{\beta})$ дисперсию σ^2 на s^2, получим оценки $\widehat{cov}(\hat{\beta})$ и $\widehat{D}(\hat{y})$:

$$\widehat{cov}(\hat{\beta}) = \widehat{K} = s^2(X^T X)^{-1}$$
$$D(\hat{y}) = x^T \widehat{K} x = s^2 x^T (X^T X)^{-1} x$$

- Так можно сделать, поскольку в формулах σ^2 или s^2 умножаются на детерминированные X-матрицу и x-вектор. (для каждой конкретной регрессионной модели x и X известны точно, без случайной составляющей)


$y_{ист}(x) = x^T \beta_{ист}$
$\hat{y}(x) = x^T \hat{\beta}$
$y_i = x_i^T \beta_{ист} + \varepsilon_i$
$\hat{y}(x_н)$
$e_и$
$y_{ист}(x_н)$
$x_н = (1, Q_н)^T$
23
22
21
20
19
18
17
16
15
4
5
7
8
9
10