# 统计学原理(Statistic)

胡华平

西北农林科技大学

经济管理学院数量经济教研室

huhuaping01@hotmail.com

2021-05-16

# 第五章 相关和回归分析

5.1 变量间关系的度量

5.2 回归分析的基本思想

5.3 OLS方法与参数估计

5.4 假设检验

5.5 拟合优度与残差分析

5.6 回归预测分析

5.7 回归报告解读

# 5.6 回归预测分析

回归预测

均值预测

个值预测

置信带

# 回归预测：引子

预测未来事件的一些惯常说法

- 算命术士：
  - “客官印堂发黑，明日必有凶象!”
- 天气预报播报词：
  - 预测西安明天是小雨，概率为95%。
  - 预测西安明天是小雨转阴，概率为95%。
  - 预测西安明天是天晴或阴天或雨天，概率为100%!
- 简要解析：
  - 人们在预测什么事件？
  - 预测多少个事件？它们发生的关系？
  - 预测如何令人信服？

# 回归预测：两类预测

一元回归模型下：

$$Y_i = \beta_1 + \beta_2 X_i + u_i$$

预测什么？

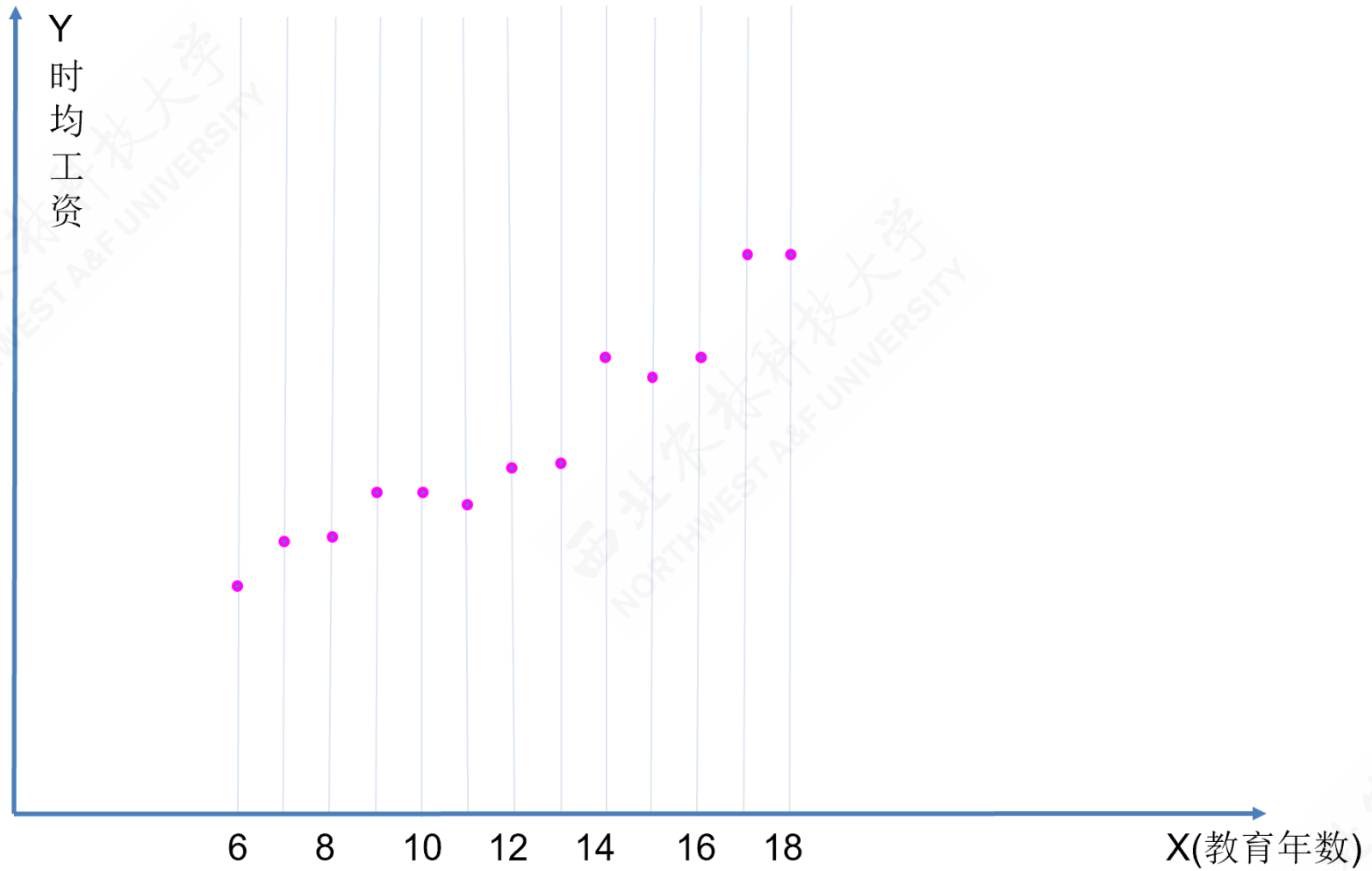
均值预测(mean prediction)：

- 给定 $X_0$，预测Y的条件均值 $E(Y|X = X_0)$

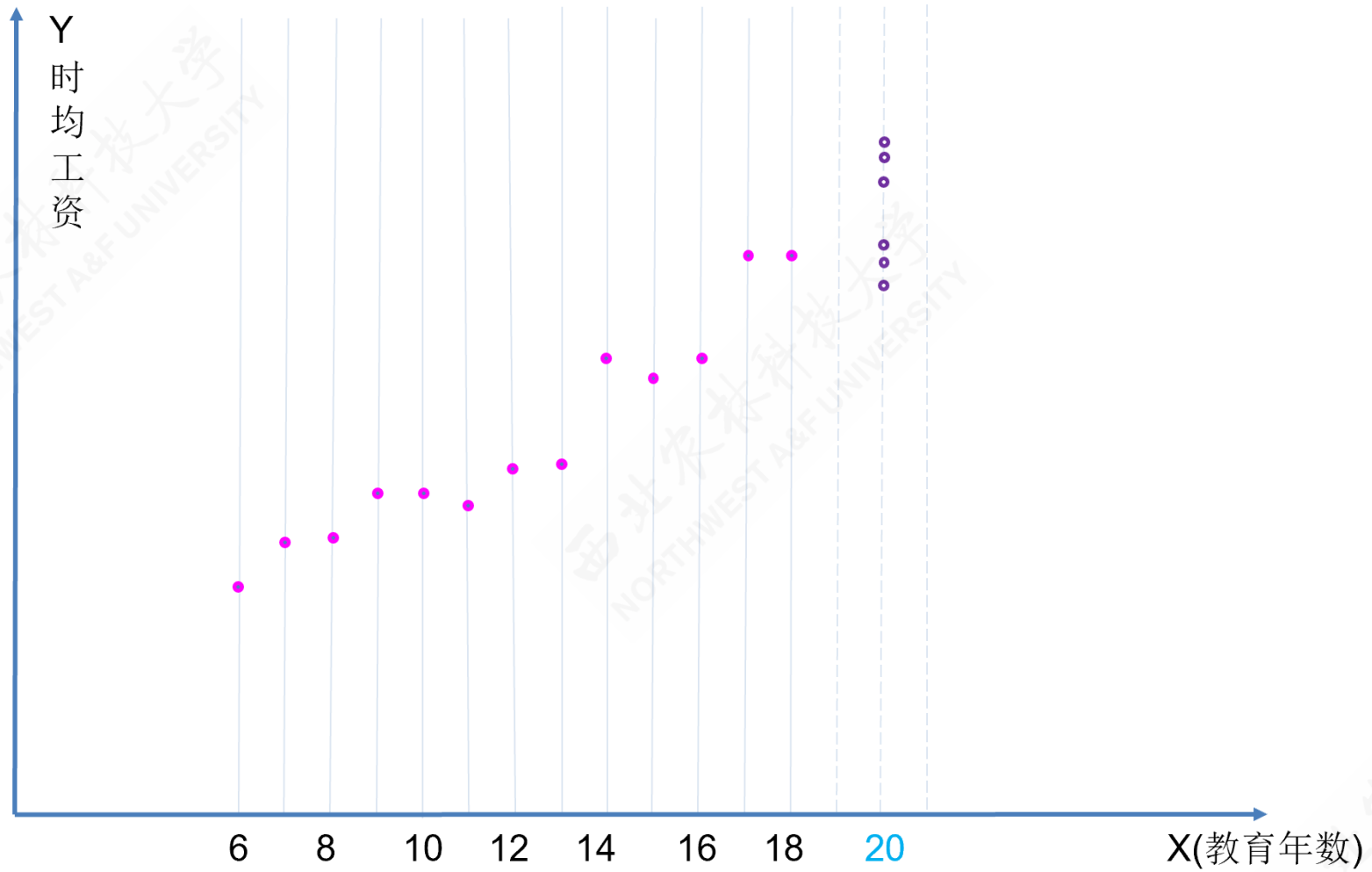
个值预测(individual prediction)：

- 给定 $X_0$，预测对应于 $X0$的Y的个别值 $(Y_0|X_0)$

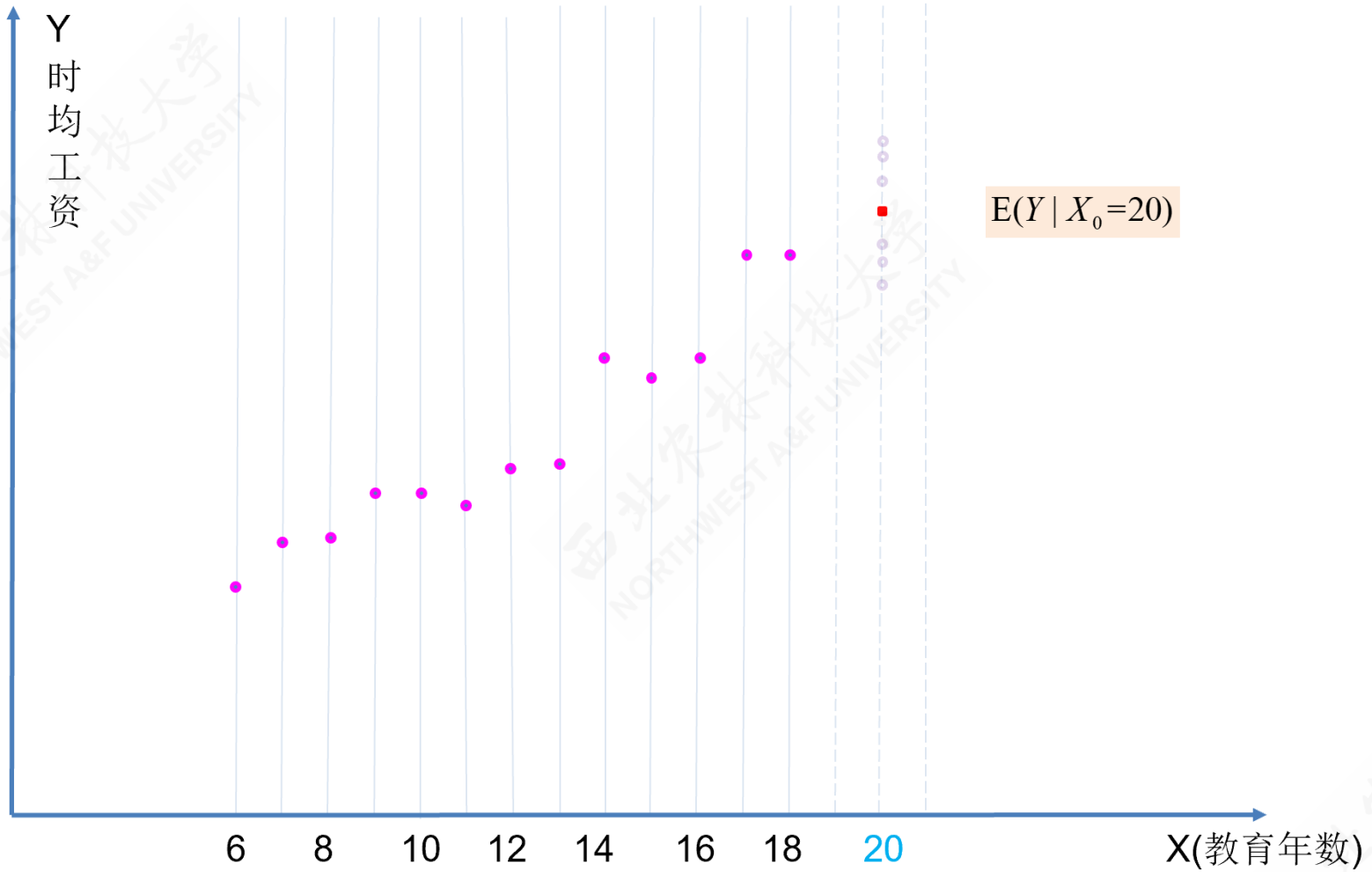
# （示例）样本内预测



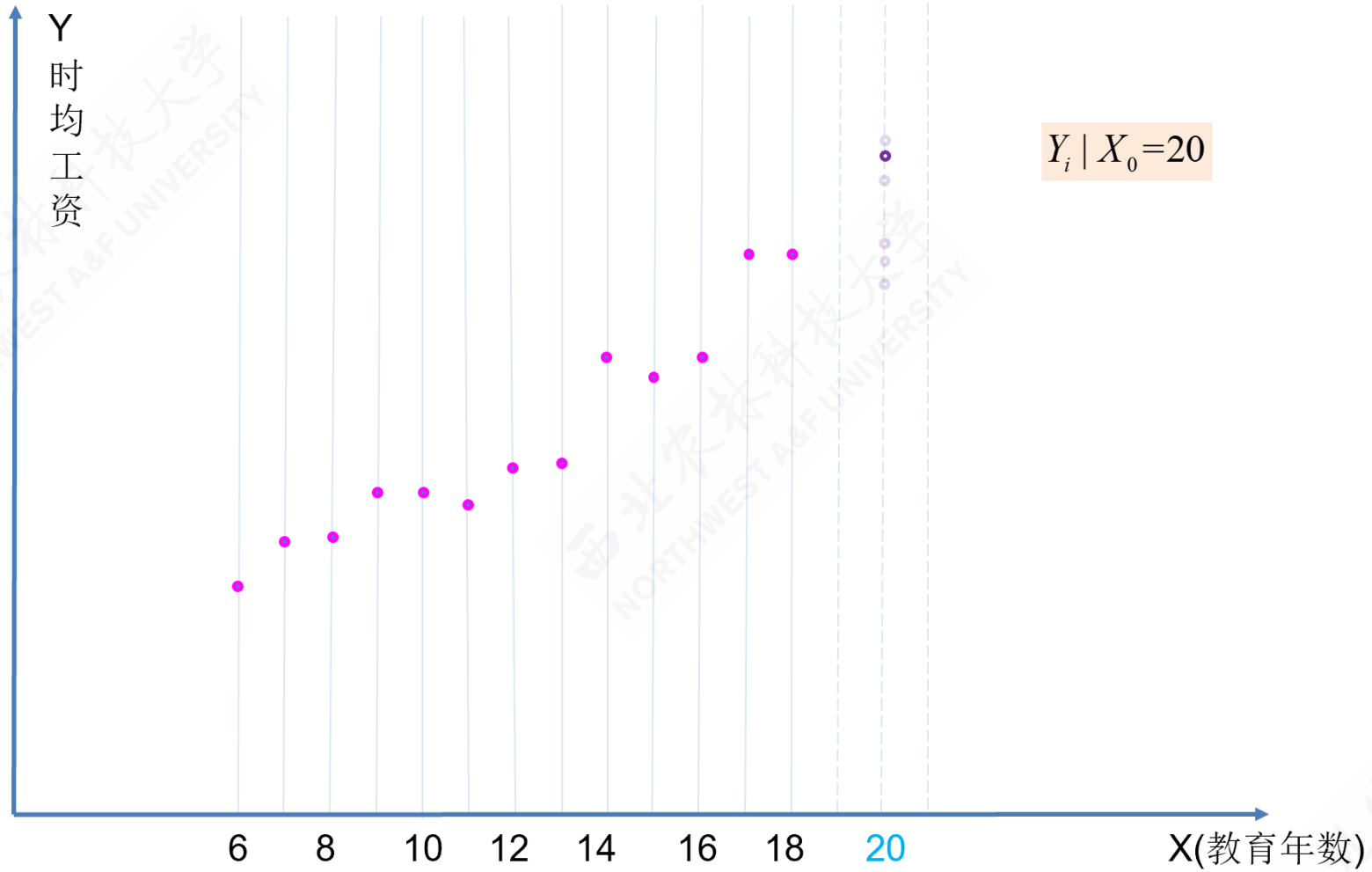
# （示例）样本外预测



# （示例）均值预测



# （示例）个值预测



# 回归预测：预测分析的关键

拿什么来预测？——样本数据？样本回归线？样本拟合值？

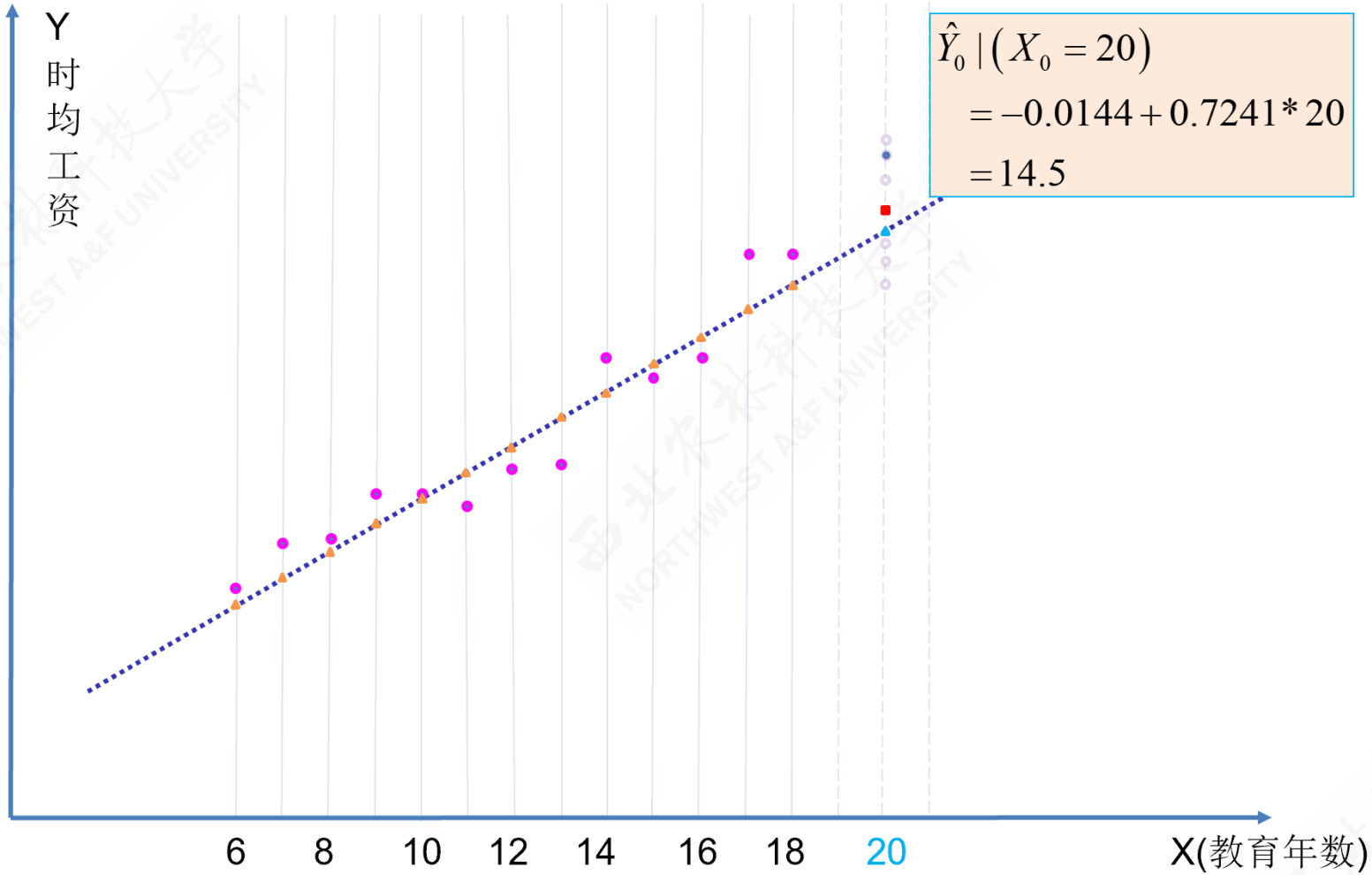
样本外拟合值 $\hat{Y}_0|X=X_0$：

- 可以证明：样本外拟合值 $\hat{Y}_0|X=X_0$是均值 $E(Y|X=X_0)$的一个BLUE
- 也可以证明：样本外拟合值 $\hat{Y}_0|X=X_0$是个值 $(Y_0|X=X_0)$的一个BLUE

工资案例中，给定 $X_0=20$，则可以得到样本外拟合值：

$$\begin{aligned}\hat{Y}_0 &= \hat{\beta}_1+\hat{\beta}_2X_0\\ &= -0.01+0.72\times 20\\ &= 14.4675\end{aligned}$$

# 回归预测：预测分析的关键



# 均值预测

在N-CLRM假设和OLS方法下，可以证明（证明过程略）给定 $X_0$下的拟合值 $\hat{Y}_0$服从如下正态分布：

$$\hat{Y}_0 \sim \mathrm{N}\left(\mu_{\hat{Y}_0}, \sigma^2_{\hat{Y}_0}\right)$$

$$\mu_{\hat{Y}_0} = E\left(\hat{Y}_0\right) = E\left(\hat{\beta}_1 + \hat{\beta}_2 X_0\right) = \beta_1 + \beta_2 X_0 = E(Y|X_0)$$

$$\mathrm{var}\left(\hat{Y}_0\right) = \sigma^2_{\hat{Y}_0} = \sigma^2 \left[\frac{1}{n} + \frac{\left(X_0 - \overline{X}\right)^2}{\sum x_i^2}\right]$$

$$\hat{Y}_0 \sim N\left(E(Y|X_0), \sigma^2 \left[\frac{1}{n} + \frac{\left(X_0 - \overline{X}\right)^2}{\sum x_i^2}\right]\right)$$

# 均值预测

对 $\hat{Y}_0$构造t统计量：

$$T=\frac{\hat{Y}_0-\mathrm{E}(\mathrm{Y}|\mathrm{X}_0)}{S_{\hat{Y}_0}}\sim t(n-2) \qquad \Leftarrow S_{\hat{Y}_0}=\sqrt{\hat{\sigma}^2\left[\frac{1}{n}+\frac{\left(X_0-\overline{X}\right)^2}{\sum x_i^2}\right]}$$

得到均值 $E(Y|X=X_0)$置信区间为：

$$\Pr\left[\hat{Y}_0-t_{1-\alpha/2}(n-2)\cdot S_{\hat{Y}_0}\le E(Y|X_0)\le \hat{Y}_0+t_{1-\alpha/2}(n-2)\cdot S_{\hat{Y}_0}\right]=1-\alpha$$

$$\Pr\left[\hat{\beta}+\hat{\beta}_2X_0-t_{1-\alpha/2}(n-2)\cdot S_{\hat{Y}_0}\le E(Y|X_0)\le \hat{\beta}+\hat{\beta}_2X_0+t_{1-\alpha/2}(n-2)\cdot S_{\hat{Y}_0}\right]=1-\alpha$$

# （案例）教育程度和时均工资：均值预测

给定 $X_0 = 20$时，根据早前计算结果：$\hat{\sigma}^2 = 0.8812$；$\bar{X} = 12.0000$；$\sum x_i^2 = 182.0000$。因此可以得到：

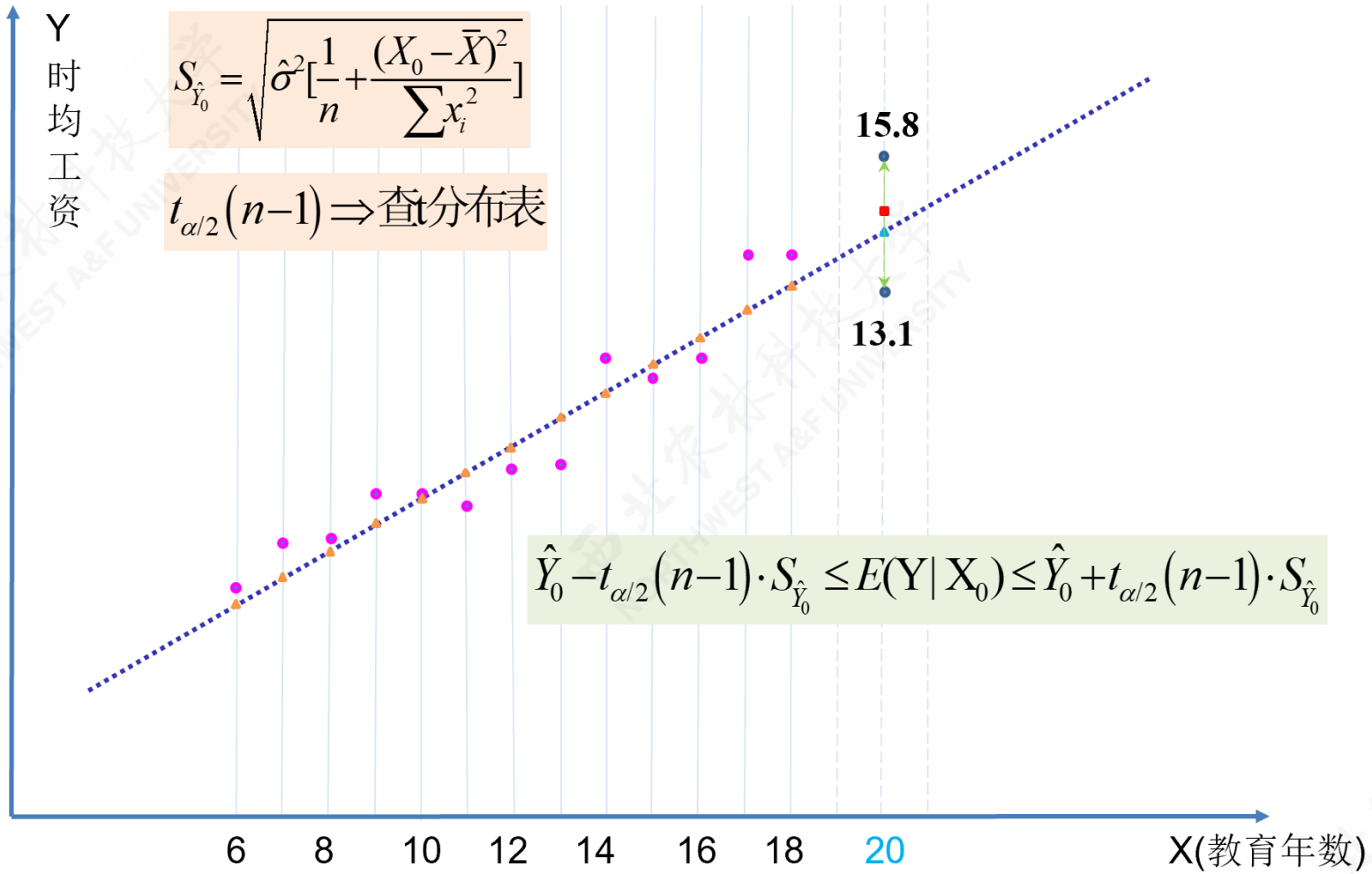
$$S_{\hat{Y}_0}^2 = \hat{\sigma}^2 \left[ \frac{1}{n} + \frac{\left(X_0 - \overline{X}\right)^2}{\sum x_i^2} \right] = 0.8812 \left( \frac{1}{13} + \frac{(20-12)^2}{182} \right) = 0.3776; \quad S_{\hat{Y}_0} = \sqrt{S_{\hat{Y}_0}^2} = 0.6145$$

$$\hat{Y}_0 = \hat{\beta}_1 + \hat{\beta}_2 X_0 = -0.0145 + 0.7241 * 20 = 14.4675$$

因此，可以计算得到均值 $E(Y|X=20)$置信区间为：

$$\begin{aligned} \hat{\beta} + \hat{\beta}_2 X_0 - t_{1-\alpha/2}(n-2) \cdot S_{\hat{Y}_0} \leq & E(Y|X_0) \leq \hat{\beta} + \hat{\beta}_2 X_0 + t_{1-\alpha/2}(n-2) \cdot S_{\hat{Y}_0} \\ 14.4675 - 1.7959 * 0.6145 \leq & E(Y|X_0 = 20) \leq 14.4675 + 1.7959 * 0.6145 \\ 13.3639 \leq & E(Y|X_0 = 20) \leq 15.5711 \end{aligned}$$

# （案例）教育程度和时均工资：均值预测



# 个值预测

在N-CLRM假设和OLS方法下，可以证明（证明过程略）给定 $X_0$下的个别值 $Y_0 = \beta_1 + \beta_2 X_0 + u_0$服从如下正态分布：

$$
\begin{aligned}
Y_0 &\sim \mathrm{N}\left(\mu_{Y_0}, \sigma^2_{Y_0}\right) \\
\mu_{Y_0} &= E\left(Y_0\right) = E\left(\beta_1 + \beta_2 X_0\right) = \beta_1 + \beta_2 X_0 \\
Var(Y_0) &= Var(u_0) = \sigma^2
\end{aligned}
$$

$$
Y_0 \sim N\left(\beta_1 + \beta_2 X_0, \sigma^2\right)
$$

# 个值预测

进一步可以构造新的随机变量 $(Y_0 - \hat{Y}_0)$，其将服从如下正态分布：

$$Y_0 \sim N\left(\beta_1 + \beta_2 X_0, \sigma^2\right)$$

$$\hat{Y}_0 \sim N\left(\beta_1 + \beta_2 X_0, \sigma^2\left[\frac{1}{n} + \frac{\left(X_0 - \overline{X}\right)^2}{\sum x_i^2}\right]\right)$$

$$Y_0 - \hat{Y}_0 \sim N\left(0, \sigma^2\left[1 + \frac{1}{n} + \frac{\left(X_0 - \overline{X}\right)^2}{\sum x_i^2}\right]\right)$$

$$Y_0 - \hat{Y}_0 \sim N\left(0, \sigma^2_{Y_0 - \hat{Y}_0}\right)$$

# 个值预测

对 $Y_0-\hat{Y}_0$构造t统计量：

$$T=\frac{(Y_0-\hat{Y}_0)}{S_{(Y_0-\hat{Y}_0)}}\sim t(n-2) \qquad \Leftarrow S_{(Y_0-\hat{Y}_0)}=\sqrt{\hat{\sigma}^2\left[1+\frac{1}{n}+\frac{\left(X_0-\overline{X}\right)^2}{\sum x_i^2}\right]}$$

得到个值 $Y_0$置信区间为：

$$\Pr\left[\hat{Y}_0-t_{1-\alpha/2}(n-2)\cdot S_{(Y_0-\hat{Y}_0)}\le Y_0\le \hat{Y}_0+t_{1-\alpha/2}(n-2)\cdot S_{(Y_0-\hat{Y}_0)}\right]=1-\alpha$$

$$\Pr\left[\hat{\beta}+\hat{\beta}_2X_0-t_{1-\alpha/2}(n-2)\cdot S_{(Y_0-\hat{Y}_0)}\le Y_0\le \hat{\beta}+\hat{\beta}_2X_0+t_{1-\alpha/2}(n-2)\cdot S_{(Y_0-\hat{Y}_0)}\right]=1-\alpha$$

给定 $X_0 = 20$时，根据早前计算结果： $\hat{\sigma}^2 = 0.8812$； $\bar{X} = 12.0000$；$\sum x_i^2 = 182.0000$。因此可以得到：

$$S^2_{(Y_0-\hat{Y}_0)} = \hat{\sigma}^2\left[1+\frac{1}{n}+\frac{\left(X_0-\overline{X}\right)^2}{\sum x_i^2}\right] = 0.8812\left(1+\frac{1}{13}+\frac{(20-12)^2}{182}\right) = 1.2588$$

$$S_{\hat{Y}_0} = \sqrt{S^2_{\hat{Y}_0}} = 1.122$$

$$\hat{Y}_0 = \hat{\beta}_1 + \hat{\beta}_2 X_0 = -0.0145 + 0.7241 * 20 = 14.4675$$

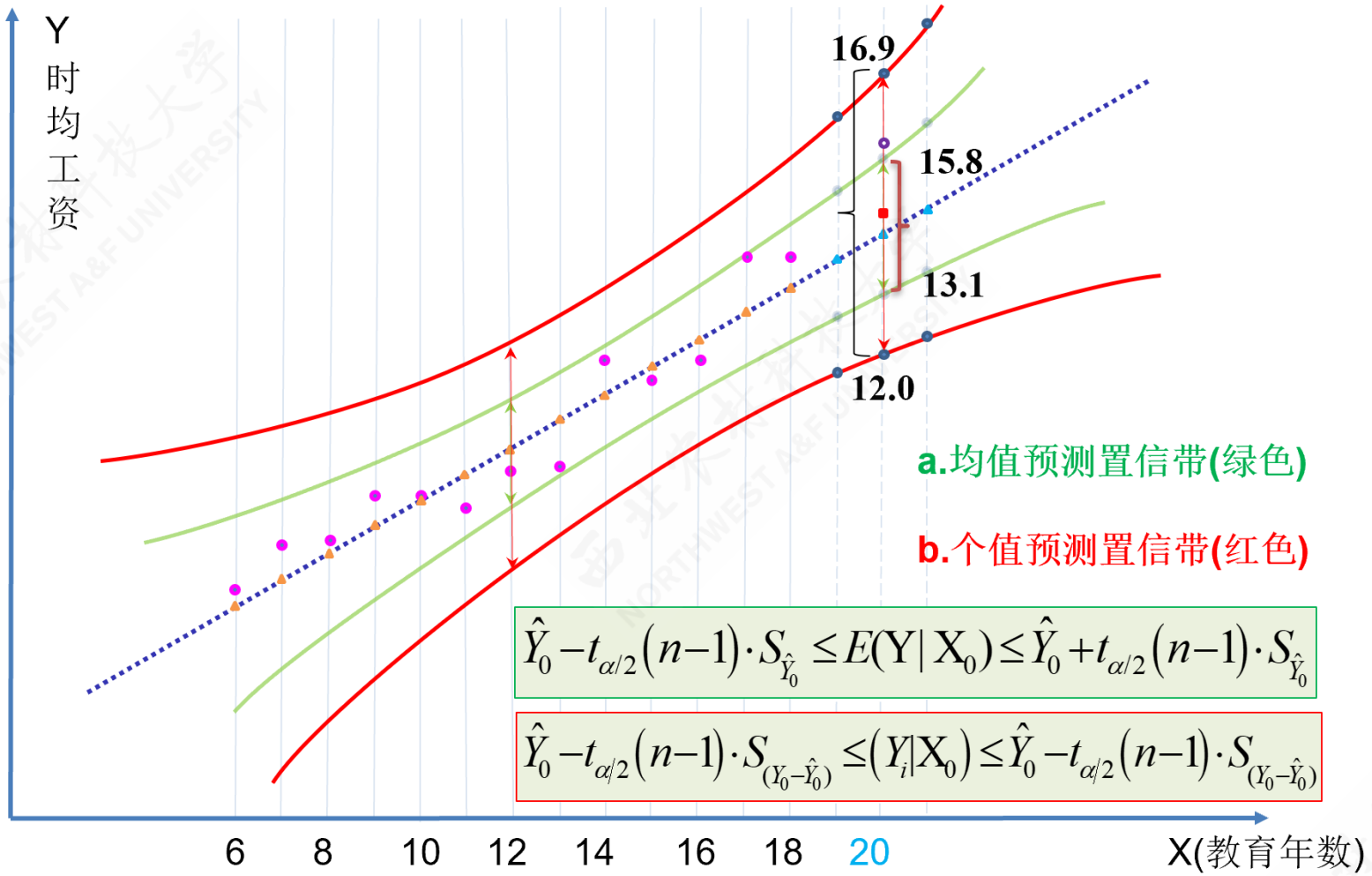
因此，可以计算得到个值 $(Y_0|X=20)$置信区间为：

$$\hat{\beta} + \hat{\beta}_2 X_0 - t_{1-\alpha/2}(n-2)\cdot S_{(Y_0-\hat{Y}_0)} \leq Y_0|X=X_0) \leq \hat{\beta} + \hat{\beta}_2 X_0 + t_{1-\alpha/2}(n-2)\cdot S_{(Y_0-\hat{Y}_0)}$$

$$14.4675 - 1.7959 * 1.122 \leq Y_0|X_0=20) \leq 14.4675 + 1.7959 * 1.122$$

$$12.4525 \leq Y_0|X_0=20) \leq 16.4824$$

# （案例）教育程度和时均工资：个值预测




a.均值预测置信带(绿色)

b.个值预测置信带(红色)

$$\hat{Y}_0 - t_{\alpha/2}(n-1)\cdot S_{\hat{Y}_0} \le E(Y|X_0) \le \hat{Y}_0 + t_{\alpha/2}(n-1)\cdot S_{\hat{Y}_0}$$

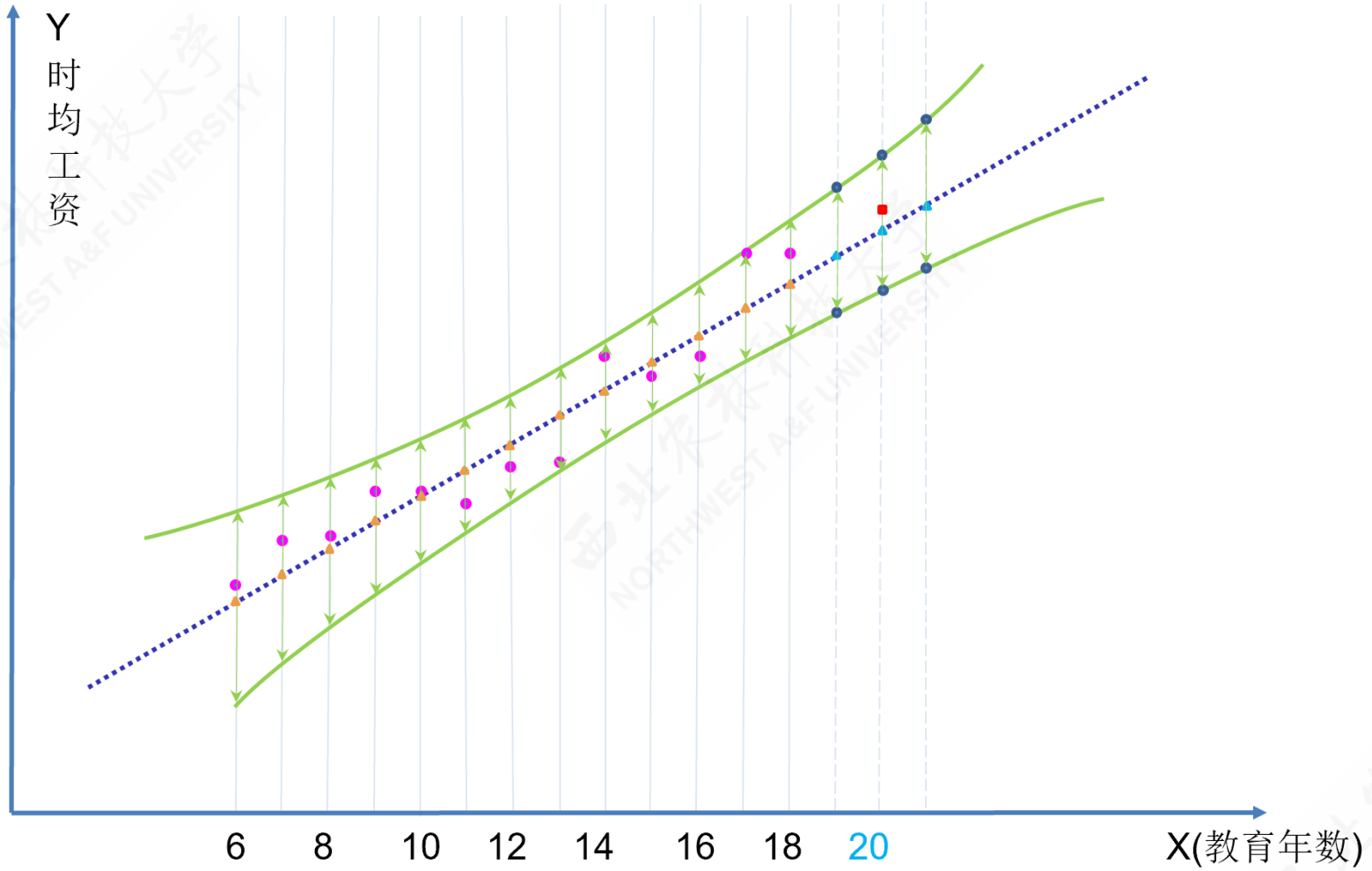
$$\hat{Y}_0 - t_{\alpha/2}(n-1)\cdot S_{(Y_0-\hat{Y}_0)} \le (Y_i|X_0) \le \hat{Y}_0 - t_{\alpha/2}(n-1)\cdot S_{(Y_0-\hat{Y}_0)}$$

# 置信带

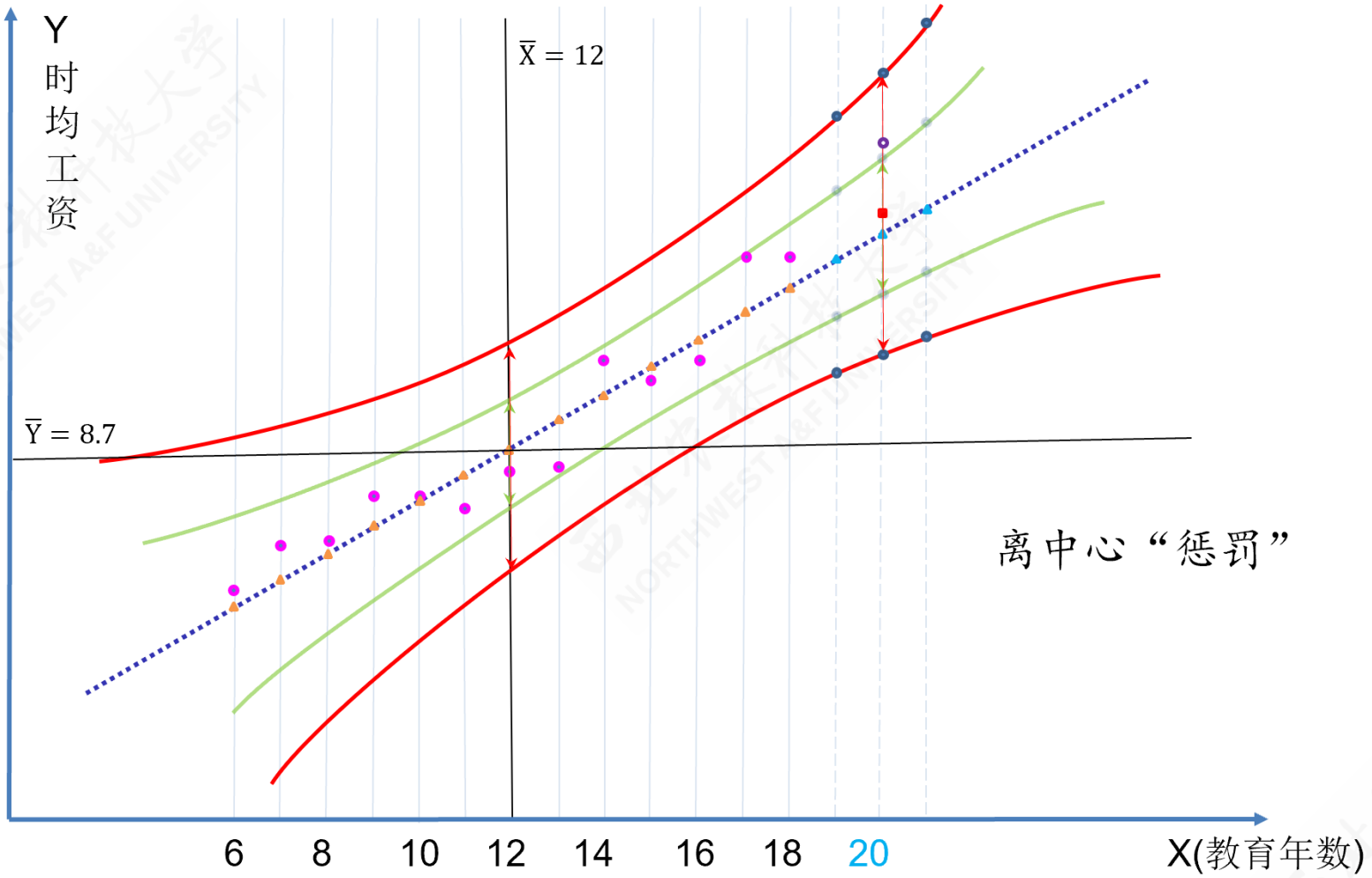
置信带(confidence interval)：对所有的X值，分别进行均值和个值分别进行预测，就能得到：

- 均值预测的置信带——总体回归函数的置信带
- 个值预测的置信带
- 预测如何可信？
  - 均值预测置信区间
  - 均值预测置信带
- 样本内置信带。——检验可靠性
- 样本外置信带。——预测未来值范围

# 置信带



# 置信带


Y
时
均
工
资
$\overline{X} = 12$
$\overline{Y} = 8.7$
离中心“惩罚”
6
8
10
12
14
16
18
20
X(教育年数)

# 置信带

如何理解置信带？

- 谁更宽？——均值预测更准确
- 何处最窄？—— 中心点 $(\bar{X}, \bar{Y}) = (12, 8.67)$是历史信息的集中代表。

# 回归预测：总结与思考

内容总结：

- 回归预测基于一套坚实严密的“底座”：OLS估计方法、CLRM假设、BLUE估计性质
- 均值预测置信带和个值预测置信带，是对预测可信度的形象表达。
- （同等条件下）均值预测比个值预测更准确（置信带宽窄）

课堂思考：

- 同样是95%置信度区间，两个人的认识是一样的么？

课后作业：工资与教育案例扩展

- 请计算置信度 $100(1-\alpha)=95\%$下， $X_0=20$时均值的置信区间。 与 $100(1-\alpha)=90\%$时相比，有什么差异？
- 99%更值得可信么？

# 本节结束