Stat 306: Finding Relationships in Data.

Lecture 5

Section 2.5 (continued)

linear regression

Age vs. Money

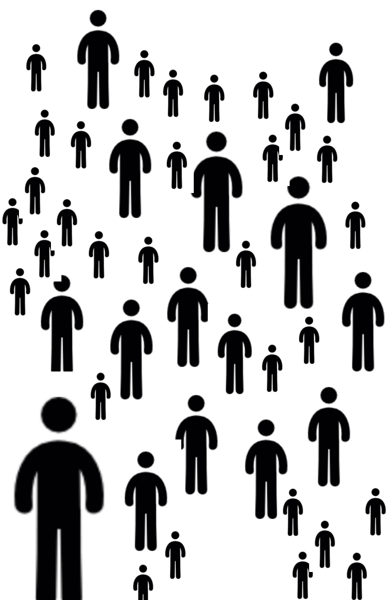
PREDICTOR *variable*

$X \longrightarrow$ Age in Years

RESPONSE *variable*

$Y \longrightarrow$ dollars ($) In bank account

Population



Population parameters

$$\beta_0, \beta_1, \sigma^2$$

Hypothesis Test

$$H_0 : \beta_1 = 0$$
$$H_1 : \beta_1 \neq 0$$

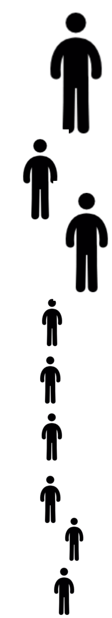
Sample statistics

b_0 = 17.7
b_1 = 0.55
s = 15.5
R^2 = 0.49

For parameter β_1 :

$$95\% \text{ C.I.} = [0.05, 1.05]$$
$$p\text{-value} = 0.036$$

Sample, n=9



x	y
82	71
45	54
71	43
22	45
29	21
9	11
12	30
18	45
24	10

Population parameter or **"something we would like to estimate"**	**Sample statistic ("estimator")**	**Estimator as a Random Variable**	**Expected Value of the estimator**	**Variance of the estimator**	**Standard Error of estimator**	**Confidence Interval**
β_0	b_0	B_0	$E[B_0]$	$Var[B_0]$	$se(b_0)$	C.I. for β_0
β_1	b_1	B_1	$E[B_1]$	$Var[B_1]$	$se(b_1)$	C.I. for β_1
σ^2	s^2	S^2	$E[S^2]$	$Var[S^2]$	$se(s^2)$	C.I. for σ^2
$\mu_Y(x)$	$(\hat{\mu}_Y(x))$	$(\hat{\mu}_Y(x))$	$E(\hat{\mu}_Y(x))$	$Var(\hat{\mu}_Y(x))$	$se(\hat{\mu}_Y(x))$	C.I. for $\mu_Y(x)$

Step 0: From θ, define estimator, $\hat{\theta}$

Step 1: Consider the sample statistic, $\hat{\theta}$, as a random variable $\hat{\Theta}$

Step 2: Determine $E[\hat{\Theta}]$ (to confirm it's unbiased) $Var[\hat{\Theta}]$ (to calculate se)

Step 3: Define se($\hat{\theta}$) = estimate of $\sqrt{\mathrm{Var}(\hat{\Theta})}$

Step 4: Define (1-α)% C.I. = $\hat{\theta} \pm c \times se(\hat{\theta})$

Population parameter or "something we would like to estimate"	**Sample statistic ("estimator")**	**Estimator as a Random Variable**	**Expected Value of the estimator**	**Variance of the estimator**	**Standard Error of estimator**	**Confidence Interval**
β_0	b_0	B_0	$E[B_0]$	$Var[B_0]$	$se(b_0)$	C.I. for β_0
β_1	b_1	B_1	$E[B_1]$	$Var[B_1]$	$se(b_1)$	C.I. for β_1
σ^2	s^2	S^2	$E[S^2]$	$Var[S^2]$	$se(s^2)$	C.I. for σ^2
$\mu_Y(x)$	$(\hat{\mu}_Y(x))$	$(\hat{\mu}_Y(x))$	$E(\hat{\mu}_Y(x))$	$Var(\hat{\mu}_Y(x))$	$se(\hat{\mu}_Y(x))$	C.I. for $\mu_Y(x)$

Population parameter or "something we would like to estimate"
β_0
β_1
σ^2
$\mu_Y(x)$

The simple linear regression model:

$$Y_i = \beta_0 + \beta_1 x_i + \epsilon_i, \quad i = 1, \ldots, n$$

where the ϵ_i's are independent normal random variables with mean 0 and variance σ^2

Therefore:

$$Y_i \sim N(\beta_0 + \beta_1 x_i, \sigma^2)$$

subpopulation mean:

$$\mathrm{E}[Y|X = x] = \beta_0 + \beta_1 x$$

Step 0:
From θ, define estimator, $\hat{\theta}$

Population parameter **or "something we would like to estimate"**	**Sample statistic ("estimator")**
β_0	b_0
β_1	b_1
σ^2	s^2
$\mu_Y(x)$	$\hat{\mu}_Y(x)$

$$b_0 = \overline{y} - b_1\overline{x}$$

$$b_1 = r_{xy}\frac{s_y}{s_x} = \sum_{i=1}^{n} a_i y_i$$, where: $a_i = \frac{(x_i - \overline{x})}{(n-1)s_x^2}$

$$s^2 = \frac{\sum_{i=1}^{n}\epsilon_i^2}{n-2}$$

$\hat{\mu}_Y(x)$ = $b_0 + b_1x = \sum_{i=1}^{n} c_i y_i$, where:

$$c_i = n^{-1} + a_i(x - \overline{x}) = n^{-1} + \frac{(x - \overline{x})(x_i - \overline{x})}{(n-1)s_x^2}$$

Step 0:
From θ, define estimator, $\hat{\theta}$

Step 1:
Consider the sample statistic, $\hat{\theta}$, as a random variable $\hat{\Theta}$

Population parameter or "something we would like to estimate"	**Sample statistic ("estimator")**	**Estimator as a Random Variable**
β_0	b_0	B_0
β_1	b_1	B_1
σ^2	s^2	S^2
$\mu_Y(x)$	$(\hat{\mu}_Y(x))$	$(\hat{\mu}_Y(x))$

$$B_0 = \bar{Y} - B_1\bar{x}$$

$$B_1 = \sum_{i=1}^{n} a_i Y_i$$, where: $a_i = \frac{(x_i - \bar{x})}{(n-1)s_x^2}$

$$S^2 = \frac{\sum_{i=1}^{n} \epsilon_i^2}{n-2}$$

$$\hat{\mu}_Y(x) = \sum_{i=1}^{n} c_i Y_i$$, where: $c_i = n^{-1} + \frac{(x-\bar{x})(x_i - \bar{x})}{(n-1)s_x^2}$

Step 0:
From θ, define estimator, $\hat{\theta}$

Step 1:
Consider the sample statistic, $\hat{\theta}$, as a random variable $\hat{\Theta}$

Population parameter or "something we would like to estimate"	**Sample statistic ("estimator")**	**Estimator as a Random Variable**
β_0	b_0	B_0
β_1	b_1	B_1
σ^2	s^2	S^2
$\mu_Y(x)$	$(\hat{\mu}_Y(x))$	$(\hat{\mu}_Y(x))$

Preview from section 2.6...

$$\hat{B}_1 \sim N\left(\beta_1, \sigma^2_{\hat{\beta}_1} = \frac{\sigma^2}{(n-1)s_x^2}\right),$$

$$\frac{\hat{B}_1 - \beta_1}{\sigma_{\hat{\beta}_1}} \sim N(0,1).$$

Step 0:
From θ, define estimator, $\hat{\theta}$

Step 1:
Consider the sample statistic, $\hat{\theta}$, as a random variable $\hat{\Theta}$

Step 2:
Determine
$E[\hat{\Theta}]$ (to confirm it's unbiased)
$Var[\hat{\Theta}]$ (to calculate se)

Population parameter or "something we would like to estimate"	**Sample statistic ("estimator")**	**Estimator as a Random Variable**	**Expected Value of the estimator**
β_0	b_0	B_0	$E[B_0]$
β_1	b_1	B_1	$E[B_1]$
σ^2	s^2	S^2	$E[S^2]$
$\mu_Y(x)$	$(\hat{\mu}_Y(x))$	$(\hat{\mu}_Y(x))$	$E(\hat{\mu}_Y(x))$

$$B_0 = \bar{Y} - B_1\bar{x}$$

$$\begin{aligned}
E[B_0] &= E[\bar{Y} - B_1\bar{X}] \\
&= \tfrac{1}{n}E[\textstyle\sum_{i=1}^{n} Y_i] - \beta_1\bar{X} \\
&= \tfrac{1}{n}\textstyle\sum_{i=1}^{n}(\beta_0 + \beta_1 X_i) - \beta_1\bar{X} \\
&= \beta_0 + \tfrac{1}{n}\textstyle\sum_{i=1}^{n}\beta_1 X_i - \beta_1\bar{X} \\
&= \beta_0 + \beta_1\tfrac{1}{n}\textstyle\sum_{i=1}^{n} X_i - \beta_1\bar{X} \\
&= \beta_0 + \beta_1\bar{X} - \beta_1\bar{X} \\
&= \beta_0
\end{aligned}$$

Therefore b_0 is "unbiased".

Step 0:
From θ, define estimator, $\hat{\theta}$

Step 1:
Consider the sample statistic, $\hat{\theta}$, as a random variable $\hat{\Theta}$

Step 2:
Determine
E[$\hat{\Theta}$] (to confirm it's unbiased)
Var[$\hat{\Theta}$] (to calculate se)

Population parameter or "something we would like to estimate"	**Sample statistic ("estimator")**	**Estimator as a Random Variable**	**Expected Value of the estimator**
β_0	b_0	B_0	$E[B_0]$ *unbiased*
β_1	b_1	B_1	$E[B_1]$
σ^2	s^2	S^2	$E[S^2]$
$\mu_Y(x)$	$(\hat{\mu}_Y(x))$	$(\hat{\mu}_Y(x))$	$E(\hat{\mu}_Y(x))$

$$B_1 = \sum_{i=1}^{n} a_i Y_i$$, where:

$$a_i = \frac{(x_i - \bar{x})}{(n-1)s_x^2}$$

$$\text{(2.56)} \quad E(\hat{B}_1) = \sum_{i=1}^{n} a_i E(Y_i) = \sum_{i=1}^{n} a_i(\beta_0 + \beta_1 x_i)$$

$$= \beta_0 \sum_{i=1}^{n} a_i + \beta_1 \sum_{i=1}^{n} a_i x_i$$

$$= 0 + \beta_1 \sum_{i=1}^{n} \frac{(x_i - \bar{x})x_i}{[(n-1)s_x^2]}$$

$$\text{(2.57)} \quad = \beta_1,$$

Therefore b_1 is "unbiased".

Step 0:
From θ, define estimator, $\hat{\theta}$

Step 1:
Consider the sample statistic, $\hat{\theta}$, as a random variable $\hat{\Theta}$

Step 2:
Determine
$E[\hat{\Theta}]$ (to confirm it's unbiased)
$Var[\hat{\Theta}]$ (to calculate se)

Population parameter or "something we would like to estimate"	**Sample statistic ("estimator")**	**Estimator as a Random Variable**	**Expected Value of the estimator**
β_0	b_0	B_0	$E[B_0]$
β_1	b_1	B_1	$E[B_1]$ *unbiased*
σ^2	s^2	S^2	$E[S^2]$
$\mu_Y(x)$	$(\hat{\mu}_Y(x))$	$(\hat{\mu}_Y(x))$	$E(\hat{\mu}_Y(x))$

$$S^2 = \frac{\sum_{i=1}^n \epsilon_i^2}{n-2}$$

$$E\{\sum_{i=1}^n (Y_i - \hat{Y}_i)^2\} = \sum_{i=1}^n E(Y_i - \hat{Y}_i)^2 = \sum_{i=1}^n Var(Y_i - \hat{Y}_i) + \{E(Y_i - \hat{Y}_i)\}^2$$
$$= \sum_{i=1}^n Var\{(Y_i - \bar{Y} - b_1(X_i - \bar{X}))^2\}$$
$$= \sum_{i=1}^n \{Var(Y_i - \bar{Y}) - 2Cov(Y_i - \bar{Y}, b_1(X_i - \bar{X})) + Var(b_1)(X_i - \bar{X})^2\}$$
$$= \sum_{i=1}^n \{Var(Y_i - \bar{Y}) - 2Cov((Y_i - \bar{Y})(X_i - \bar{X}), b_1) + Var(b_1)(X_i - \bar{X})^2\}$$
$$= \sum_{i=1}^n \{Var(\varepsilon_i - \bar{\varepsilon}) - 2Cov((Y_i - \bar{Y})(X_i - \bar{X}), b_1) + Var(b_1)(X_i - \bar{X})^2\}$$
$$= (n-1)\sigma^2 - 2Cov(\sum_{i=1}^n (Y_i - \bar{Y})(X_i - \bar{X}), b_1) + Var(b_1)\sum_{i=1}^n (X_i - \bar{X})^2$$
$$= (n-1)\sigma^2 - 2Cov(b_1 \sum_{i=1}^n (X_i - \bar{X})^2, b_1) + Var(b_1)\sum_{i=1}^n (X_i - \bar{X})^2$$
$$= (n-1)\sigma^2 - Var(b_1)\sum_{i=1}^n (X_i - \bar{X})^2 = (n-2)\sigma^2.$$

$$E[\sum_{i=1}^n \epsilon^2] = (n-2)\sigma^2$$

$$E[\frac{\sum_{i=1}^n \epsilon^2}{(n-2)}] = \sigma^2$$

$$E[S^2] = \sigma^2$$

Therefore s^2 is "unbiased".

credit: https://web.njit.edu/~wguo/Math644_2012/Math644_Chapter%201_part2.pdf

Step 0:
From θ, define estimator, $\hat{\theta}$

Step 1:
Consider the sample statistic, $\hat{\theta}$, as a random variable $\hat{\Theta}$

Step 2:
Determine
$E[\hat{\Theta}]$ (to confirm it's unbiased)
$Var[\hat{\Theta}]$ (to calculate se)

Population parameter or "something we would like to estimate"	**Sample statistic ("estimator")**	**Estimator as a Random Variable**	**Expected Value of the estimator**
β_0	b_0	B_0	$E[B_0]$ *unbiased*
β_1	b_1	B_1	$E[B_1]$ *unbiased*
σ^2	s^2	S^2	$E[S^2]$ *unbiased*
$\mu_Y(x)$	$(\hat{\mu}_Y(x))$	$(\hat{\mu}_Y(x))$	$E(\hat{\mu}_Y(x))$

$$\hat{\mu}_Y(x) = \sum_{i=1}^{n} c_i Y_i \quad , \text{ where:}$$

$$c_i = n^{-1} + \frac{(x - \overline{x})(x_i - \overline{x})}{(n-1)s_x^2}$$

$$E\left[\hat{\mu}_Y(x)\right] = \sum_{i=1}^{n} c_i E(Y_i) = \sum_{i=1}^{n} c_i(\beta_0 + \beta_1 x_i)$$

(2.60)

$$= \beta_0 \sum_{i=1}^{n} c_i + \beta_1 \sum_{i=1}^{n} c_i x_i$$

(2.61)

(2.62)

$$= \beta_0 + \beta_1\overline{x} + \beta_1(x - \overline{x}) \sum_{i=1}^{n} a_i x_i$$

$$= \beta_0 + \beta_1\overline{x} + \beta_1(x - \overline{x})$$

$$= \beta_0 + \beta_1 x$$

Therefore the subpopulation mean" is "unbiased".

Step 0:
From θ, define estimator, $\hat{\theta}$

Step 1:
Consider the sample statistic, $\hat{\theta}$, as a random variable $\hat{\Theta}$

Step 2:
Determine
$E[\hat{\Theta}]$ (to confirm it's unbiased)
$Var[\hat{\Theta}]$ (to calculate se)

Population parameter **or "something we would like to estimate"**	**Sample statistic ("estimator")**	**Estimator as a Random Variable**	**Expected Value of the estimator**	**Variance of the estimator**
β_0	b_0	B_0	$E[B_0]$ *unbiased*	$Var[B_0]$
β_1	b_1	B_1	$E[B_1]$ *unbiased*	$Var[B_1]$
σ^2	s^2	S^2	$E[S^2]$ *unbiased*	$Var[S^2]$
$\mu_Y(x)$	$(\hat{\mu}_Y(x))$	$(\hat{\mu}_Y(x))$	$E(\hat{\mu}_Y(x))$ *unbiased*	$Var(\hat{\mu}_Y(x))$

Step 0:
From θ, define estimator, $\hat{\theta}$

Step 2:
Determine
$E[\hat{\Theta}]$ (to confirm it's unbiased)
$Var[\hat{\Theta}]$ (to calculate se)

Population parameter or "something we would like to estimate"	**Sample statistic ("estimator")**	**Expected Value of the estimator**	**Variance of the estimator**
β_0	b_0	$E[B_0]$ *unbiased*	$Var[B_0]$
β_1	b_1	$E[B_1]$ *unbiased*	$Var[B_1]$
σ^2	s^2	$E[S^2]$ *unbiased*	$Var[S^2]$
$\mu_Y(x)$	$(\hat{\mu}_Y(x))$	E $(\hat{\mu}_Y(x))$ *unbiased*	Var $(\hat{\mu}_Y(x))$

$$B_0 = \bar{Y} - B_1\bar{x}$$

$$Var(B_0) = Var(\textstyle\sum_{i=1}^{n} Y_i/n) + \bar{X}^2 Var(B_1)$$

$$= \sigma^2\left(\frac{1}{n} + \frac{\bar{X}^2}{\sum_{i=1}^{n}(X_i - \bar{X})^2}\right)$$

Step 0:
From θ, define estimator, $\hat{\theta}$

Step 2:
Determine
$E[\hat{\Theta}]$ (to confirm it's unbiased)
$Var[\hat{\Theta}]$ (to calculate se)

Population parameter or "something we would like to estimate"	Sample statistic ("estimator")	Expected Value of the estimator	Variance of the estimator
β_0	b_0	$E[B_0]$ *unbiased*	$Var[B_0]$
β_1	b_1	$E[B_1]$ *unbiased*	$Var[B_1]$
σ^2	s^2	$E[S^2]$ *unbiased*	$Var[S^2]$
$\mu_Y(x)$	$(\hat{\mu}_Y(x))$	$E(\hat{\mu}_Y(x))$ *unbiased*	$Var(\hat{\mu}_Y(x))$

$B_1 = \sum_{i=1}^{n} a_i Y_i$, where:

$$a_i = \frac{(x_i - \bar{x})}{(n-1)s_x^2}$$

$$\begin{aligned} Var(\hat{B}_1) &= \sum_{i=1}^{n} a_i^2 Var(Y_i) \\ &= \sigma^2 \sum_{i=1}^{n} a_i^2 \\ &= \sigma^2 \sum_{i=1}^{n} \frac{(x_i - \overline{x})^2}{[(n-1)s_x^2]^2} \\ &= \frac{\sigma^2}{(n-1)s_x^2} \end{aligned}$$

Step 0:
From θ, define estimator, $\hat{\theta}$

Step 2:
Determine
$E[\hat{\Theta}]$ (to confirm it's unbiased)
$Var[\hat{\Theta}]$ (to calculate se)

Population parameter **or "something we would like to estimate"**	**Sample statistic ("estimator")**	**Expected Value of the estimator**	**Variance of the estimator**
β_0	b_0	$E[B_0]$ *unbiased*	$Var[B_0]$
β_1	b_1	$E[B_1]$ *unbiased*	$Var[B_1]$
σ^2	s^2	$E[S^2]$ *unbiased*	$Var[S^2]$
$\mu_Y(x)$	$(\hat{\mu}_Y(x))$	$E\left(\hat{\mu}_Y(x)\right)$ *unbiased*	$Var\left(\hat{\mu}_Y(x)\right)$

We will skip this....

Step 0:
From θ, define estimator, $\hat{\theta}$

Step 2:
Determine
$E[\hat{\Theta}]$ (to confirm it's unbiased)
$Var[\hat{\Theta}]$ (to calculate se)

Population parameter or "something we would like to estimate"	**Sample statistic ("estimator")**	**Expected Value of the estimator**	**Variance of the estimator**
β_0	b_0	$E[B_0]$ *unbiased*	$Var[B_0]$
β_1	b_1	$E[B_1]$ *unbiased*	$Var[B_1]$
σ^2	s^2	$E[S^2]$ *unbiased*	$Var[S^2]$
$\mu_Y(x)$	$(\hat{\mu}_Y(x))$	$E(\hat{\mu}_Y(x))$ *unbiased*	$Var\,(\hat{\mu}_Y(x))$

$\hat{\mu}_Y(x) = \sum_{i=1}^{n} c_i Y_i$, where:

$$c_i = n^{-1} + \frac{(x-\overline{x})(x_i-\overline{x})}{(n-1)s_x^2}$$

$$(2.63) \qquad \mathrm{Var}\left[\hat{\mu}_Y(x)\right] = \sum_{i=1}^{n} c_i^2 \mathrm{Var}\left(Y_i\right)$$

$$= \sigma^2 \sum_{i=1}^{n} c_i^2$$

$$= \sigma^2 \sum_{i=1}^{n} \left\{n^{-1} + (x-\overline{x})(x_i-\overline{x})/[(n-1)s_x^2]\right\}^2$$

$$= \sigma^2 \sum_{i=1}^{n} \left\{n^{-2} + \frac{(x-\overline{x})^2(x_i-\overline{x})^2}{[(n-1)s_x^2]^2} + 2n^{-1}\frac{(x-\overline{x})(x_i-\overline{x})}{[(n-1)s_x^2]}\right\}$$

$$= \sigma^2 \left\{n^{-1} + \frac{(x-\overline{x})^2 \sum_i (x_i-\overline{x})^2}{[(n-1)s_x^2]^2} + 0\right\}$$

$$(2.66) \qquad = \sigma^2 \left\{n^{-1} + \frac{(x-\overline{x})^2}{[(n-1)s_x^2]}\right\}$$

Step 3:
Define
se($\hat{\theta}$) =
estimate of $\sqrt{\mathrm{Var}\,(\hat{\Theta})}$

Step 4:
Define
(1-α)% C.I. =
$\hat{\theta} \pm c \times se(\hat{\theta})$

$$Var(B_0) = \sigma^2\left(\frac{1}{n} + \frac{\bar{X}^2}{\sum_{i=1}^n (X_i - \bar{X})^2}\right)$$

$$\Rightarrow se(b_0) = s\sqrt{\left(\frac{1}{n} + \frac{\bar{X}^2}{\sum_{i=1}^n (X_i - \bar{X})^2}\right)}$$

Variance of the estimator	**Standard Error of estimator**	**Confidence Interval**
$Var[B_0]$	$se(b_0)$	C.I. for β_0
$Var[B_1]$	$se(b_1)$	C.I. for β_1
$Var[S^2]$	$se(s^2)$	C.I. for σ^2
Var $(\hat{\mu}_Y(x))$	$se(\hat{\mu}_Y(x))$	C.I. for $\mu_Y(x)$

$$\mathrm{Var}\,(\hat{B}_1) = \frac{\sigma^2}{(n-1)s_x^2}$$

$$\Rightarrow \quad se(\hat{\beta}_1) = \frac{\hat{\sigma}}{\sqrt{n-1}\,s_x} \qquad (2.46)$$

where:

$$\hat{\sigma} = s = \sqrt{\frac{\sum_{i=1}^{n} e_i^2}{n-2}}$$

Step 3:
Define
se($\hat{\theta}$) =
estimate of $\sqrt{\mathrm{Var}\,(\hat{\Theta})}$

Step 4:
Define
(1-α)% C.I. =
$\hat{\theta} \pm c \times se(\hat{\theta})$

Variance of the estimator	**Standard Error of estimator**	**Confidence Interval**
$\mathrm{Var}[B_0]$	$se(b_0)$	C.I. for β_0
$\mathrm{Var}[B_1]$	$se(b_1)$	C.I. for β_1
$\mathrm{Var}[S^2]$	$se(s^2)$	C.I. for σ^2
Var $(\hat{\mu}_Y(x))$	$se(\hat{\mu}_Y(x))$	C.I. for $\mu_Y(x)$

Step 3:
Define
se($\hat{\theta}$) =
estimate of $\sqrt{\text{Var}(\hat{\Theta})}$

Step 4:
Define
(1-α)% C.I. =
$\hat{\theta} \pm c \times se(\hat{\theta})$

Variance of the estimator	Standard Error of estimator	Confidence Interval
$Var[B_0]$	$se(b_0)$	C.I. for β_0
$Var[B_1]$	$se(b_1)$	C.I. for β_1
$Var[S^2]$	$se(s^2)$	C.I. for σ^2
Var $(\hat{\mu}_Y(x))$	$se(\hat{\mu}_Y(x))$	C.I. for $\mu_Y(x)$

$\Rightarrow$ We will skip this….

$$\text{Var}\left[\hat{\mu}_Y(x)\right] = \sigma^2\left\{n^{-1} + \frac{(x-\overline{x})^2}{[(n-1)s_x^2]}\right\}$$

$$\Rightarrow se(\hat{\mu}_Y(x)) = \hat{\sigma} \times \sqrt{\frac{1}{n} + \frac{(x-\overline{x})^2}{(n-1)s_x^2}} \qquad (2.47)$$

where:

$$\hat{\sigma} = s = \sqrt{\frac{\sum_{i=1}^{n} e_i^2}{n-2}}$$

Step 3:
Define
se($\hat{\theta}$) =
estimate of $\sqrt{\text{Var}\,(\hat{\Theta})}$

Step 4:
Define
(1-α)% C.I. =
$\hat{\theta} \pm c \times se(\hat{\theta})$

Variance of the estimator	Standard Error of estimator	Confidence Interval
$\text{Var}[B_0]$	$se(b_0)$	C.I. for β_0
$\text{Var}[B_1]$	$se(b_1)$	C.I. for β_1
$\text{Var}[S^2]$	$se(s^2)$	C.I. for σ^2
Var $(\hat{\mu}_Y(x))$	$se(\hat{\mu}_Y(x))$	C.I. for $\mu_Y(x)$

Step 4:
Define
(1-α)% C.I. =
$\hat{\theta} \pm c \times se(\hat{\theta})$

A confidence interval for a parameter θ commonly has the form

$$\hat{\theta} \pm c \times se(\hat{\theta}),$$

$$b_0 = \overline{y} - b_1\overline{x}$$

$$se(b_0) = s\sqrt{\left(\frac{1}{n} + \frac{\bar{X}^2}{\sum_{i=1}^{n}(X_i - \bar{X})^2}\right)}$$

95% C.I. for $\beta_0 =$

$$\left[\ \bar{y} - b_1\bar{X} - t_{n-2,0.975} \cdot s\sqrt{\left(\frac{1}{n} + \frac{\bar{X}^2}{\sum_{i=1}^{n}(X_i - \bar{X})^2}\right)}\ ,\ \bar{y} - b_1\bar{X} + t_{n-2,0.975} \cdot s\sqrt{\left(\frac{1}{n} + \frac{\bar{X}^2}{\sum_{i=1}^{n}(X_i - \bar{X})^2}\right)}\ \right]$$

Standard Error of estimator	**Confidence Interval**
$se(b_0)$	C.I. for β_0
$se(b_1)$	C.I. for β_1
$se(s^2)$	C.I. for σ^2
$se(\hat{\mu}_Y(x))$	C.I. for $\mu_Y(x)$

Step 4:
Define
(1-α)% C.I. =
$\hat{\theta} \pm c \times se(\hat{\theta})$

Standard Error of estimator	**Confidence Interval**
se(b_0)	C.I. for β_0
se(b_1)	C.I. for β_1
se(s^2)	C.I. for σ^2
$se(\hat{\mu}_Y(x))$	C.I. for $\mu_Y(x)$

$$se(\hat{\beta}_1) = \frac{\hat{\sigma}}{\sqrt{n-1}\, s_x}$$

Then we have :

95% C.I. For β_1 =

$$[b_1 - t_{n-2,0.975}\frac{\hat{\sigma}}{\sqrt{n-1}s_x}, \quad b_1 + t_{n-2,0.975}\frac{\hat{\sigma}}{\sqrt{n-1}s_x}]$$

where: $$b_1 = r_{xy}\frac{s_y}{s_x}$$

$$\hat{\sigma} = s = \sqrt{\frac{\sum_{i=1}^{n} e_i^2}{n-2}}$$

Step 4:
Define
(1-α)% C.I. =
$\hat{\theta} \pm c \times se(\hat{\theta})$

Standard Error of estimator	**Confidence Interval**
$se(b_0)$	C.I. for β_0
$se(b_1)$	C.I. for β_1
$se(s^2)$	C.I. for σ^2
$se(\hat{\mu}_Y(x))$	C.I. for $\mu_Y(x)$

We will skip this….

Step 4:
Define
(1-α)% C.I. =
$\hat{\theta} \pm c \times se(\hat{\theta})$

Standard Error of estimator	**Confidence Interval**
$se(b_0)$	C.I. for β_0
$se(b_1)$	C.I. for β_1
$se(s^2)$	C.I. for σ^2
$se(\hat{\mu}_Y(x))$	C.I. for $\mu_Y(x)$

The 95% confidence interval for subpopulation mean $\mu_Y(x) = \beta_0 + \beta_1 x$ is

$$\hat{\mu}_Y(x) \pm t_{n-2,0.975} \times se(\hat{\mu}_Y(x)), \qquad (2.43)$$

where:

$$se(\hat{\mu}_Y(x)) = \hat{\sigma} \times \sqrt{\frac{1}{n} + \frac{(x-\overline{x})^2}{(n-1)s_x^2}}$$

and:

$$\hat{\mu}_Y(x) = \hat{\beta}_0 + \hat{\beta}_1 x.$$

and:

$$\hat{\sigma} = s = \sqrt{\frac{\sum_{i=1}^n e_i^2}{n-2}}$$

Step 0: From θ, define estimator, $\hat{\theta}$

Step 1: Consider the sample statistic, $\hat{\theta}$, as a random variable $\hat{\Theta}$

Step 2: Determine $E[\hat{\Theta}]$ (to confirm it's unbiased) $Var[\hat{\Theta}]$ (to calculate se)

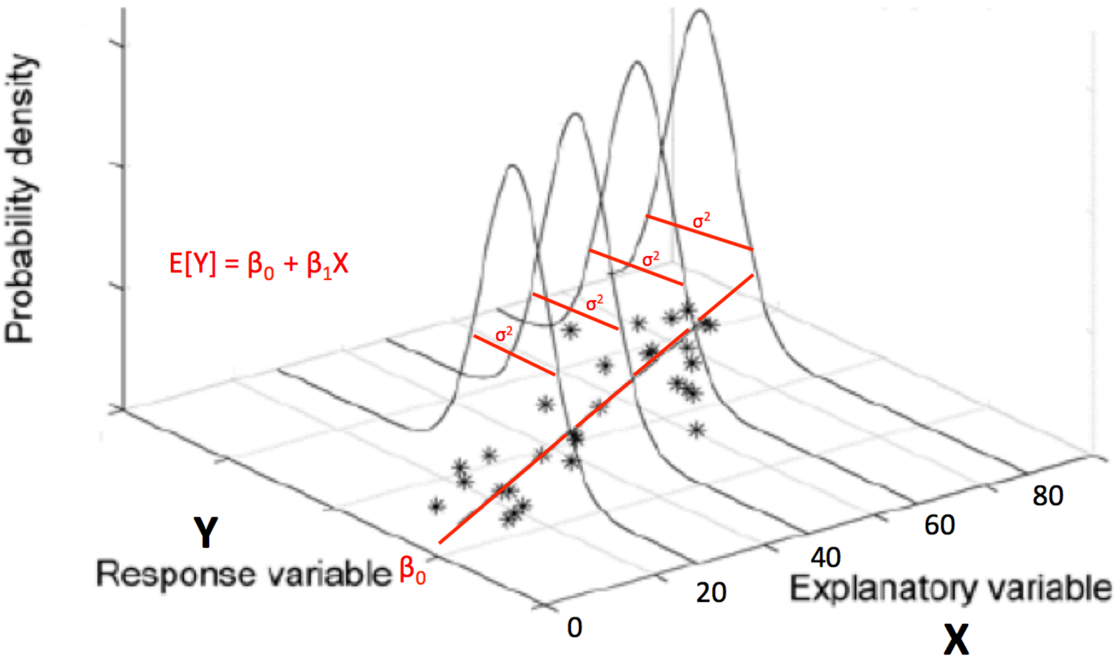
Step 3: Define se($\hat{\theta}$) = estimate of $\sqrt{\text{Var}(\hat{\Theta})}$

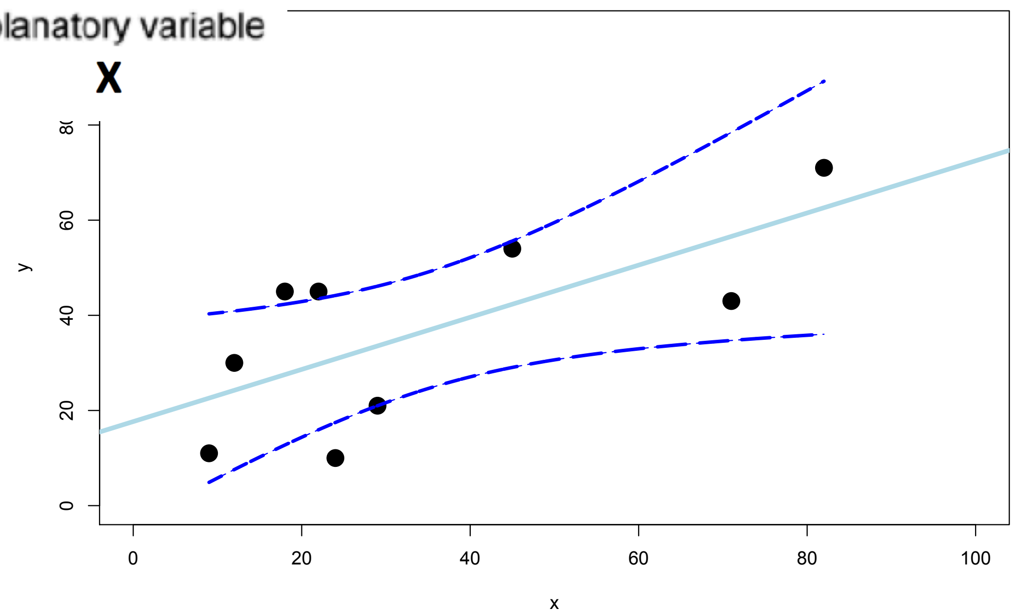
Step 4: Define (1-α)% C.I. = $\hat{\theta} \pm c \times se(\hat{\theta})$

Population parameter or "something we would like to estimate"	**Sample statistic ("estimator")**	**Estimator as a Random Variable**	**Expected Value of the estimator**	**Variance of the estimator**	**Standard Error of estimator**	**Confidence Interval**
β_0	b_0	B_0	$E[B_0]$	$Var[B_0]$	$se(b_0)$	C.I. for β_0
β_1	b_1	B_1	$E[B_1]$	$Var[B_1]$	$se(b_1)$	C.I. for β_1
σ^2	s^2	S^2	$E[S^2]$	$Var[S^2]$	$se(s^2)$	C.I. for σ^2
$\mu_Y(x)$	$(\hat{\mu}_Y(x))$	$(\hat{\mu}_Y(x))$	$E(\hat{\mu}_Y(x))$	$Var(\hat{\mu}_Y(x))$	$se(\hat{\mu}_Y(x))$	C.I. for $\mu_Y(x)$

- Questions?

- Confused about homogeneity vs. non-consistent width of confidence intervals?



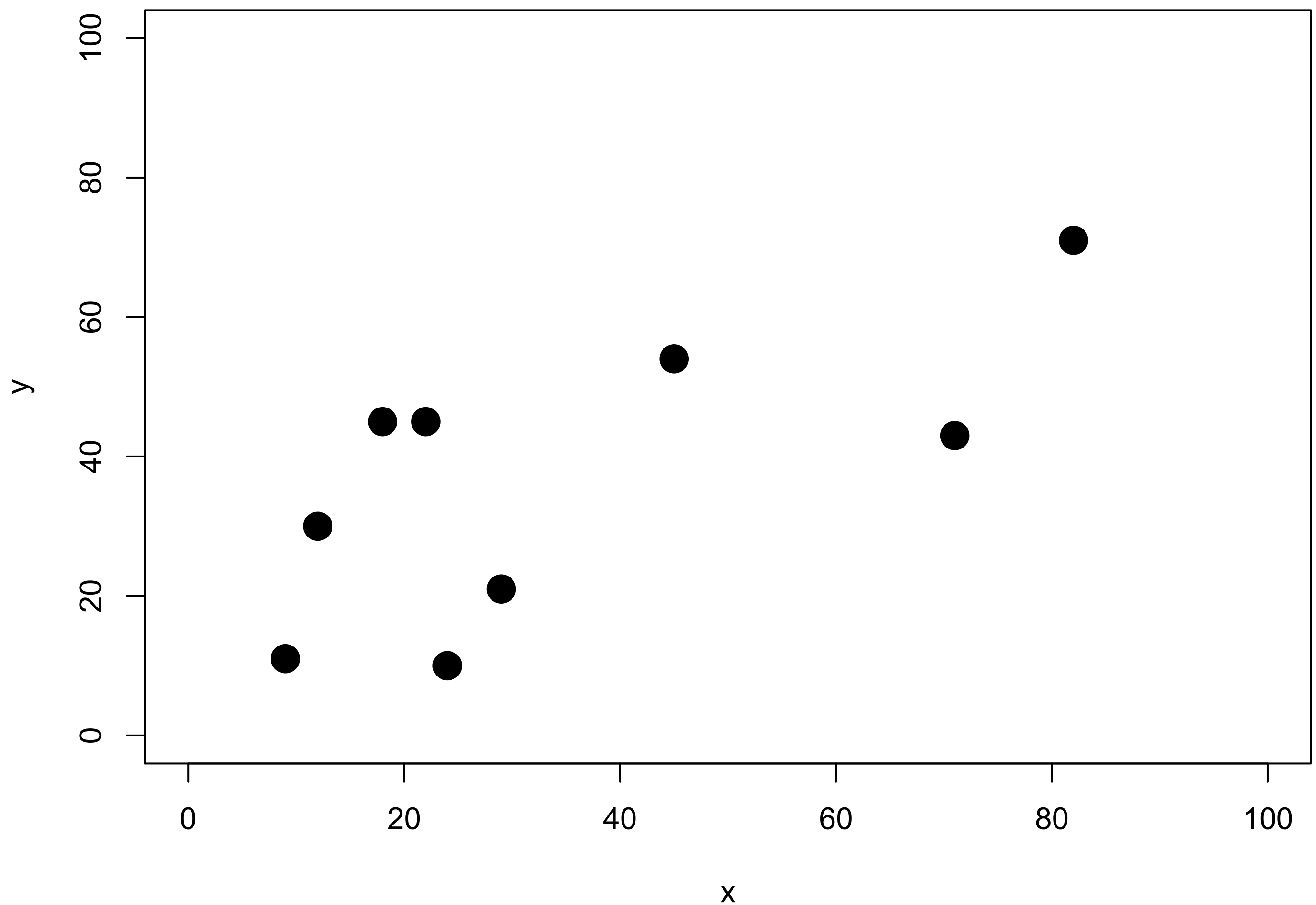


- Confused about homogeneity vs. non-consistent width of confidence intervals?

```
> ############################################################
> # Linear regression example
> ############################################################
> # x and n are fixed values
> x <- c(82, 45, 71, 22, 29, 9, 12, 18, 24)
> n <- 9
>
> # y is a realization of the random variable "Y", i.e. "observed data":
> y <- c(71, 54, 43, 45, 21, 11, 30, 45, 10)
>
> plot(y~x, xlim=c(0,100), ylim=c(0,100), pch=20, cex=3)
>
> # beta0, beta1, and sigma2 are population parameters
> # let's pretend that we know the values of these parameters:
> beta0 <- 20
> beta1 <- 0.5
> sigma2 <- 100
> |
```

- Confused about homogeneity vs. non-consistent width of confidence intervals?

We have plotted the "observed data" (i.e. one realization of the random vector **Y**):

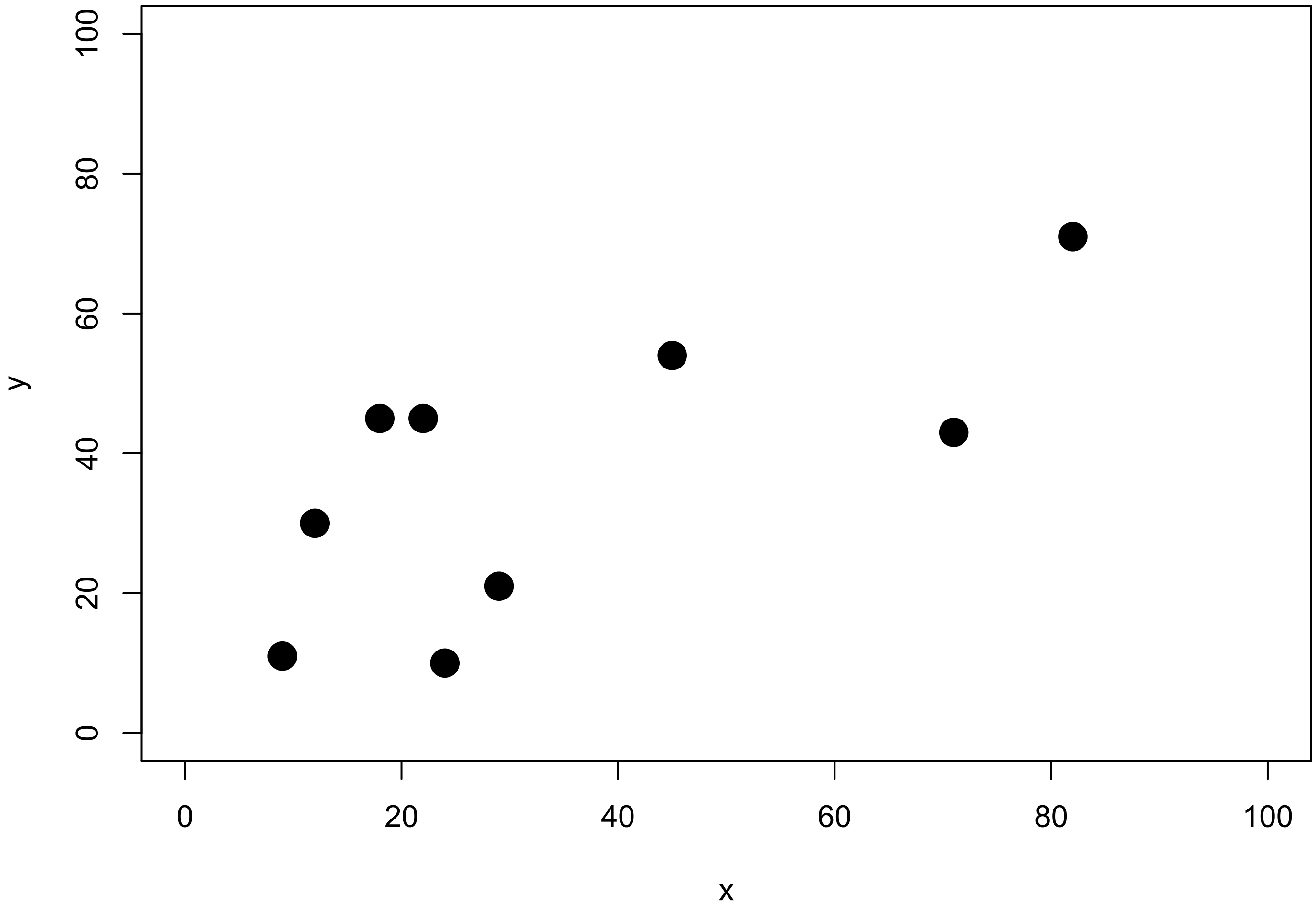


- Confused about homogeneity vs. non-consistent width of confidence intervals?

```
> ############################################################
> # Now we introduce the random variables:
>
> # epsilon (unknown) is a random variable
> epsilon <- rnorm(n, mean=0, sd=sqrt(sigma2))
>
> # Y (unknown) is a random variable
> Y <- beta0 + beta1*x + epsilon
>
> # Sample statistics (also known as "estimators")
> # can be considered as random variables:
>
> # sample means:
> xbar <- (1/n)*sum(x)
> ybar <- (1/n)*sum(Y)
>
> # sample standard deviations:
> sx <- sqrt( sum((x-xbar)^2)/(n-1) )
> sy <- sqrt( sum((Y-ybar)^2)/(n-1) )
>
> # sample covariance and sample correlation:
> sxy <- (1/(n-1))*sum((x-xbar)*(Y-ybar))
> rxy <- sxy/(sx*sy)
>
> # best estimators for beta0 and beta1 parameters
> beta1hat <- rxy*sy/sx
> beta0hat <- ybar-beta1hat*xbar
```

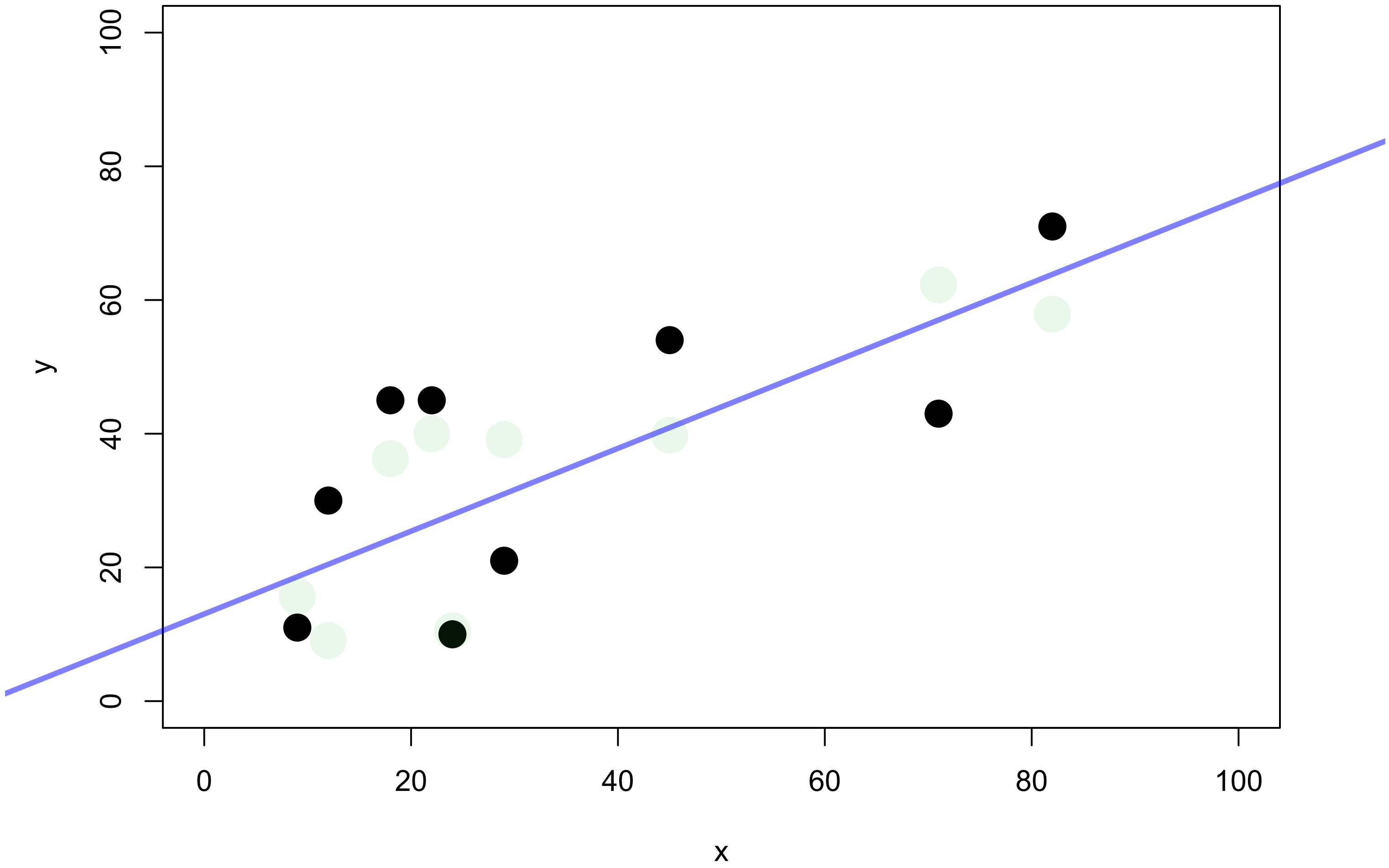
- Confused about homogeneity vs. non-consistent width of confidence intervals?

```
> residuals <- y - beta0hat - beta1hat*x
> s <- sqrt( (1/(n-2))*sum(residuals^2))
```



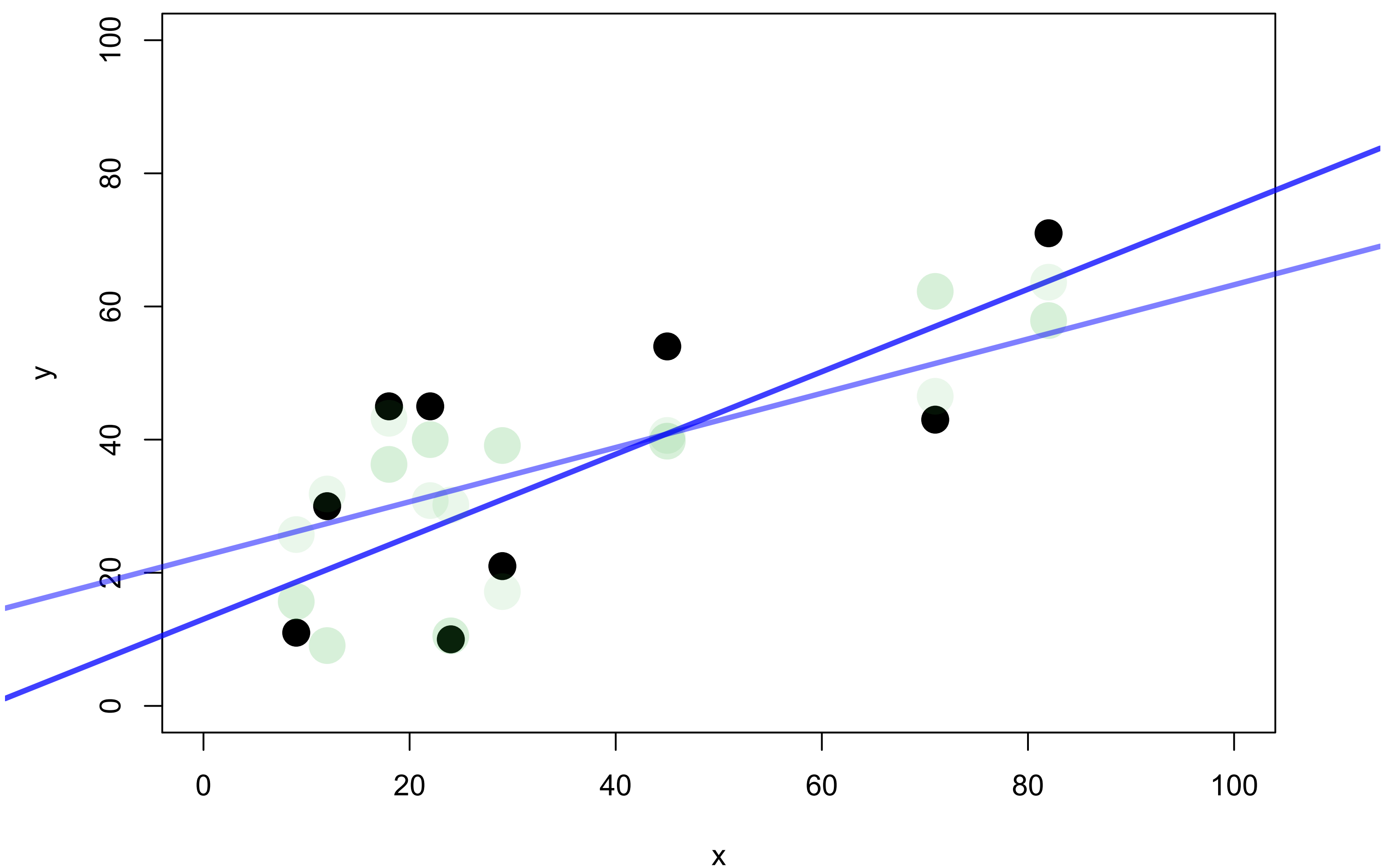
- Confused about homogeneity vs. non-consistent width of confidence intervals?

```
> # Let's plot another a realization of the random variable "Y"
>
> points(x,Y,pch=20,cex=4, col=rgb(0.2,0.7,0.25,alpha=0.10))
> abline(beta0hat, beta1hat, col=rgb(0,0,1,alpha=0.50), lwd=3)
```



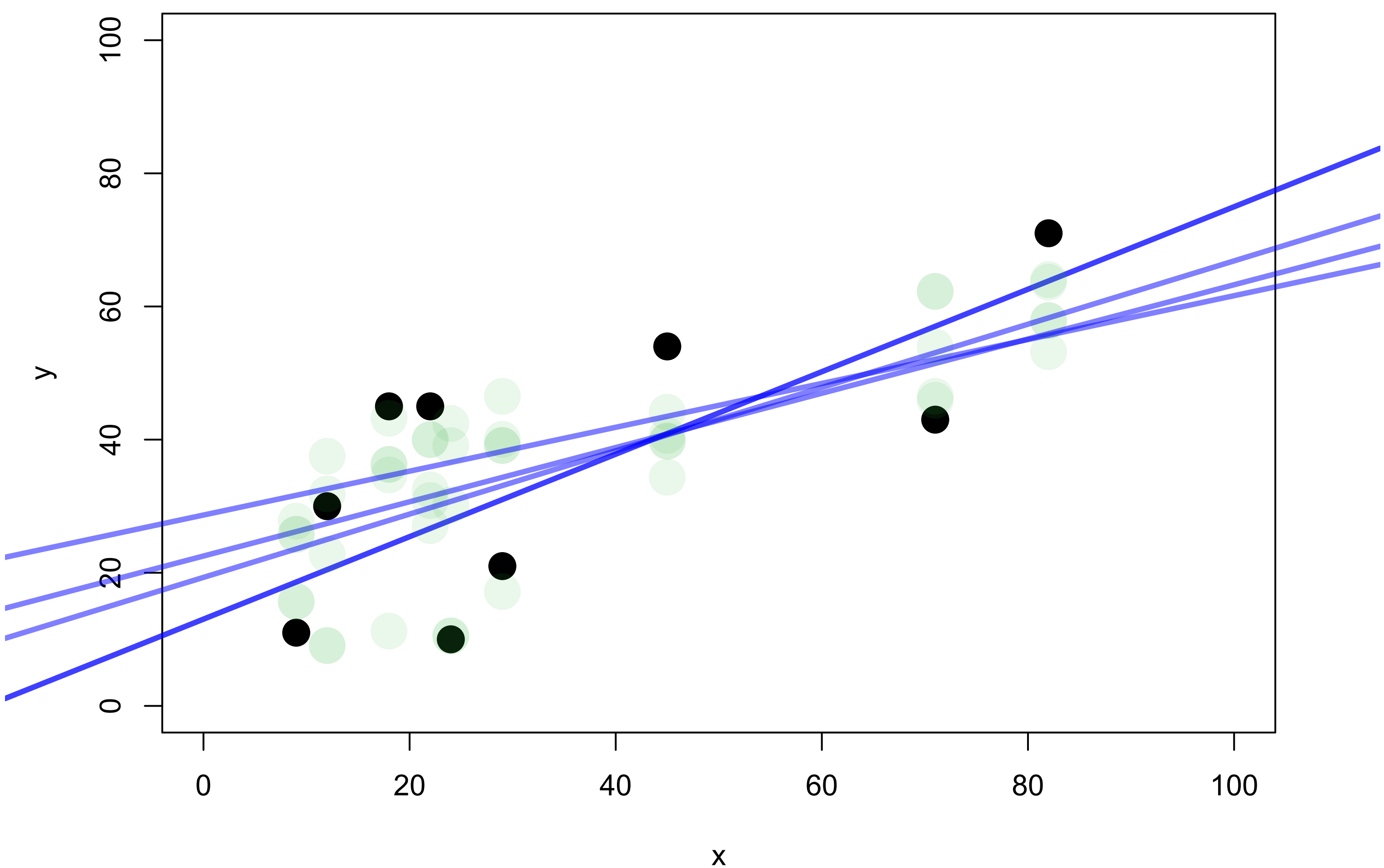
- Confused about homogeneity vs. non-consistent width of confidence intervals?

```
> # Let's plot another a realization of the random variable "Y"
>
> points(x,Y,pch=20,cex=4, col=rgb(0.2,0.7,0.25,alpha=0.10))
> abline(beta0hat, beta1hat, col=rgb(0,0,1,alpha=0.50), lwd=3)
```



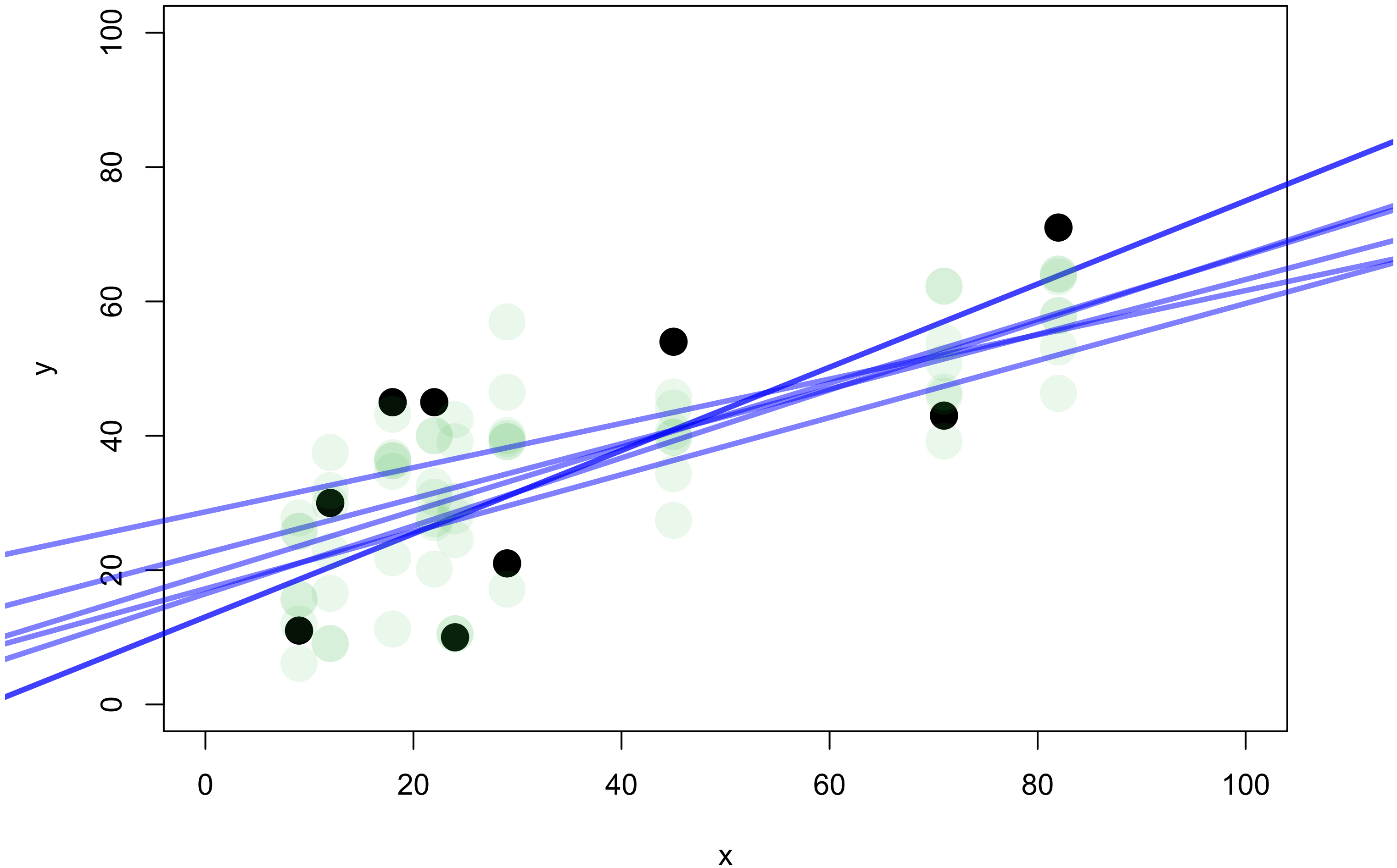
- Confused about homogeneity vs. non-consistent width of confidence intervals?

```
> # Let's plot another a realization of the random variable "Y"
>
> points(x,Y,pch=20,cex=4, col=rgb(0.2,0.7,0.25,alpha=0.10))
> abline(beta0hat, beta1hat, col=rgb(0,0,1,alpha=0.50), lwd=3)
```



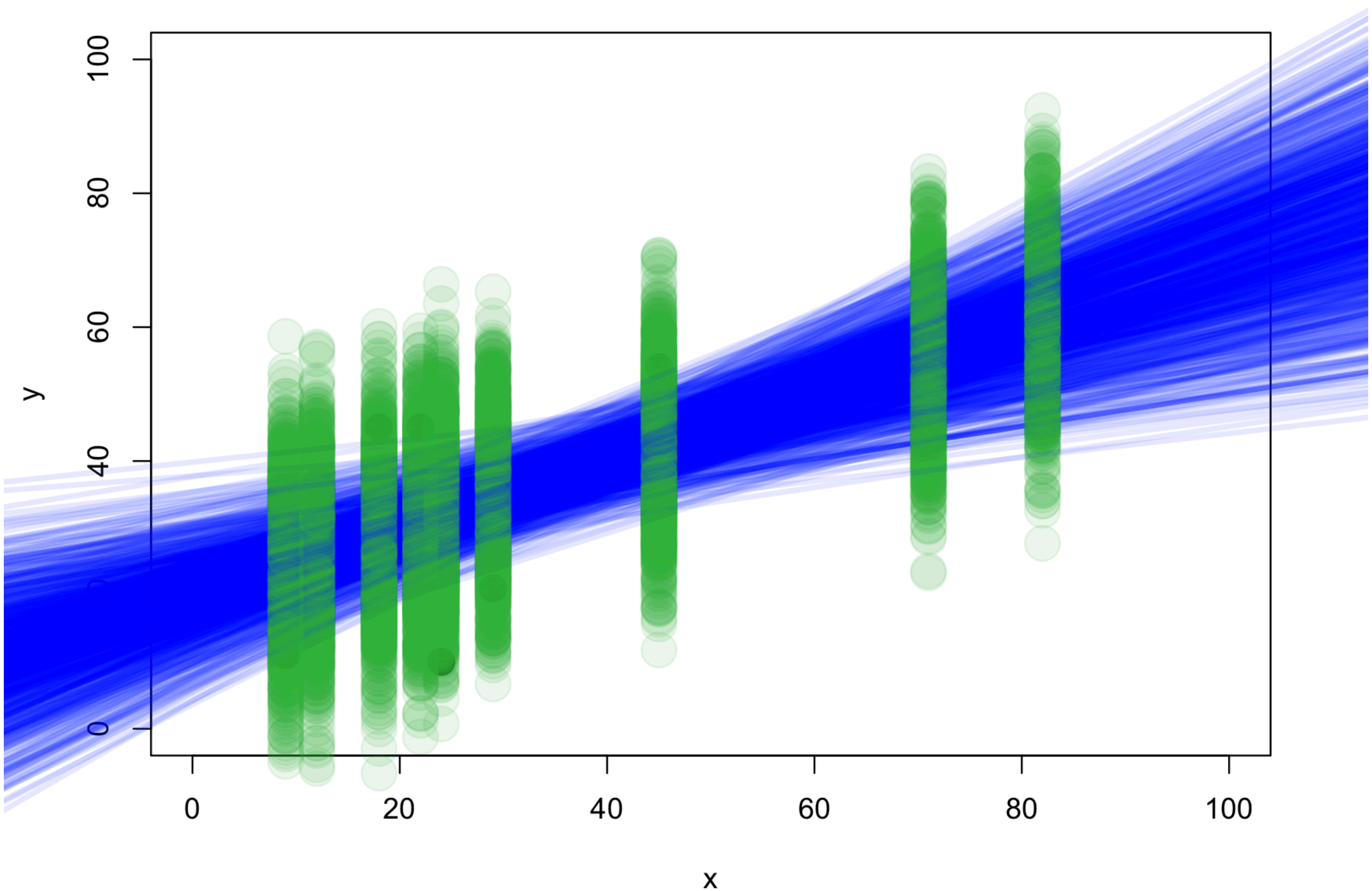
- Confused about homogeneity vs. non-consistent width of confidence intervals?

```
> # Let's plot another a realization of the random variable "Y"
>
> points(x,Y,pch=20,cex=4, col=rgb(0.2,0.7,0.25,alpha=0.10))
> abline(beta0hat, beta1hat, col=rgb(0,0,1,alpha=0.50), lwd=3)
```

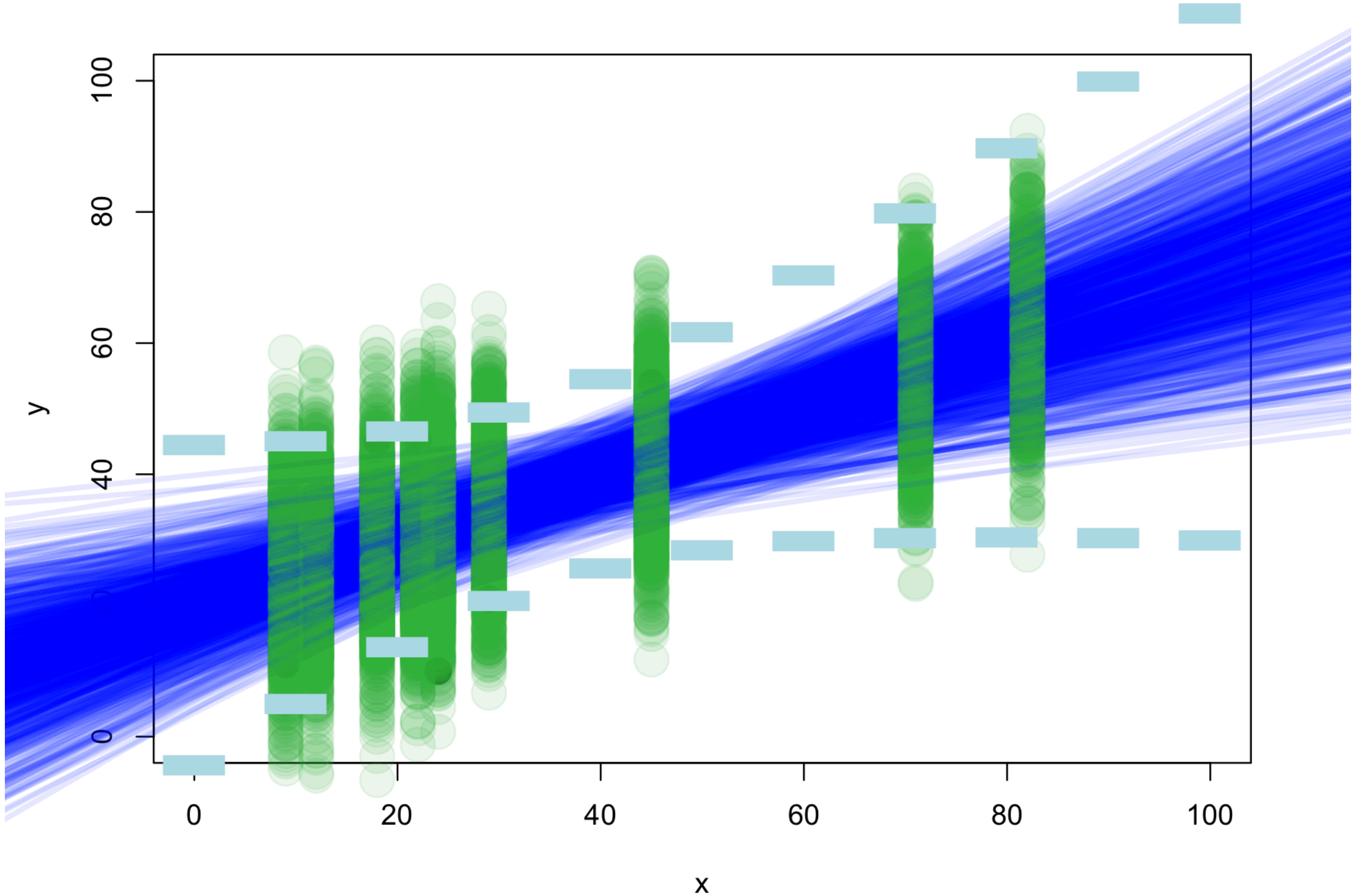


- Confused about homogeneity vs. non-consistent width of confidence intervals?

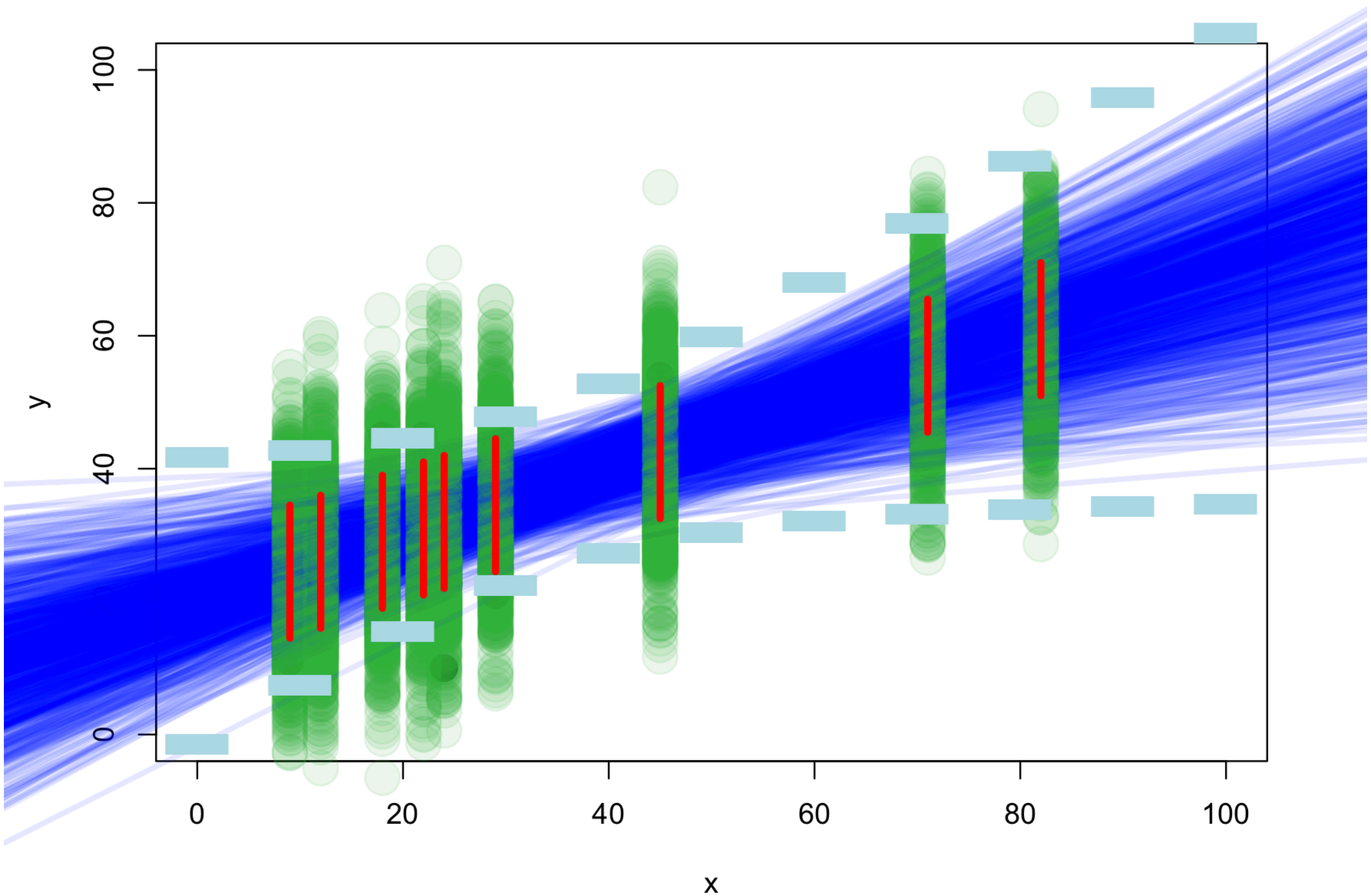
```
> # Let's plot another a realization of the random variable "Y"
>
> points(x,Y,pch=20,cex=4, col=rgb(0.2,0.7,0.25,alpha=0.10))
> abline(beta0hat, beta1hat, col=rgb(0,0,1,alpha=0.50), lwd=3)
```



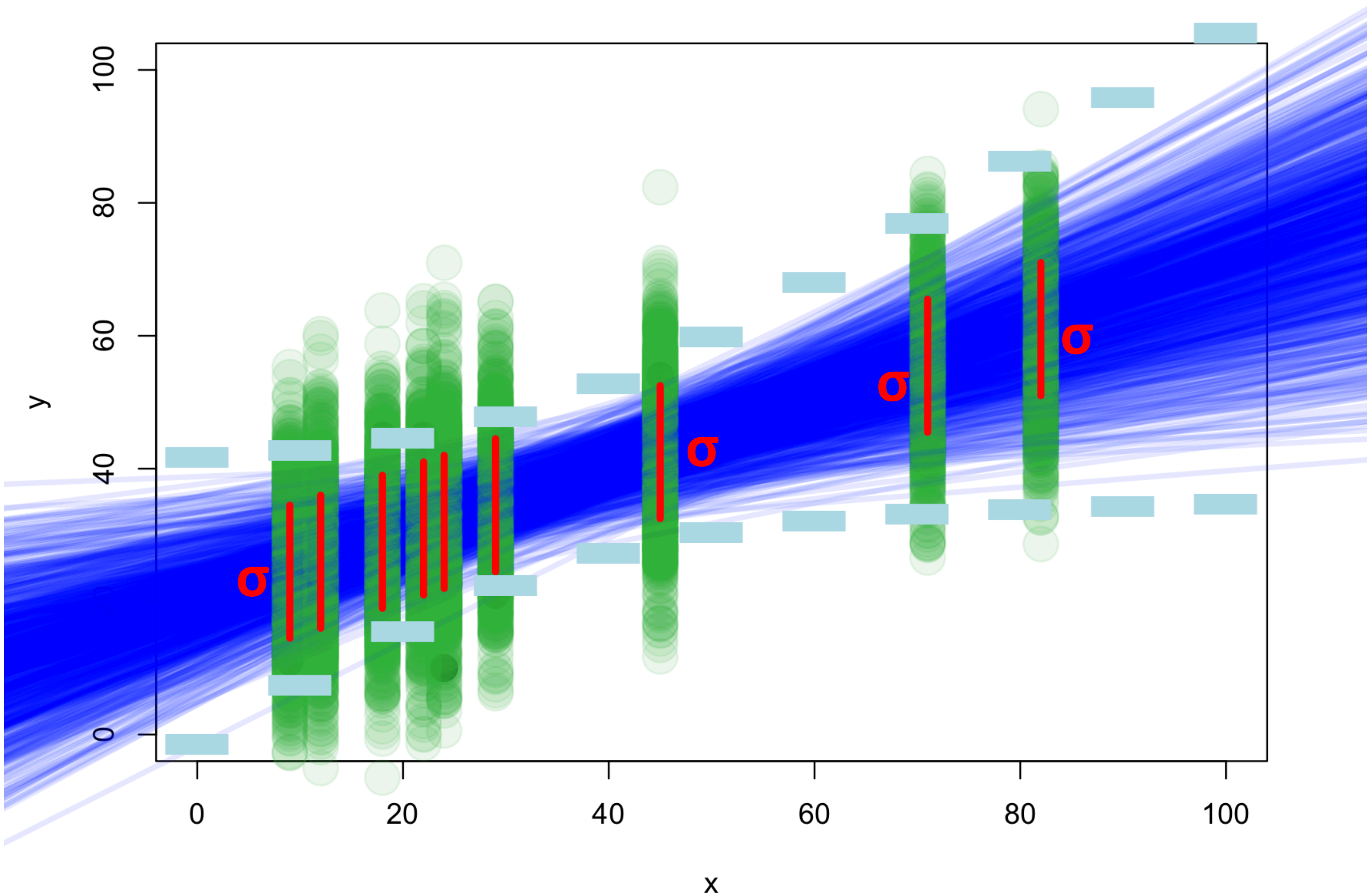
```
> # plot the 95% confidence interval for a series of subpopulation means:
> # this should look like a confidence interval for the regr
> for(myx in c(0,10,20,30,40,50,60,70,80,90,100)){
+ muhat_x <- beta0+beta1*myx
+ muhat_x
+ lowerCI <- muhat_x - qt(0.975,n-2) * s * sqrt(1/n + ((myx-xbar)^2)/((n-1)*sx^2))
+ upperCI <-  muhat_x + qt(0.975,n-2) * s * sqrt(1/n + ((myx-xbar)^2)/((n-1)*sx^2))
+
+ points(myx, lowerCI, pch="-", cex=8, col="lightblue")
+ points(myx, upperCI, pch="-", cex=8, col="lightblue")
+ }
```



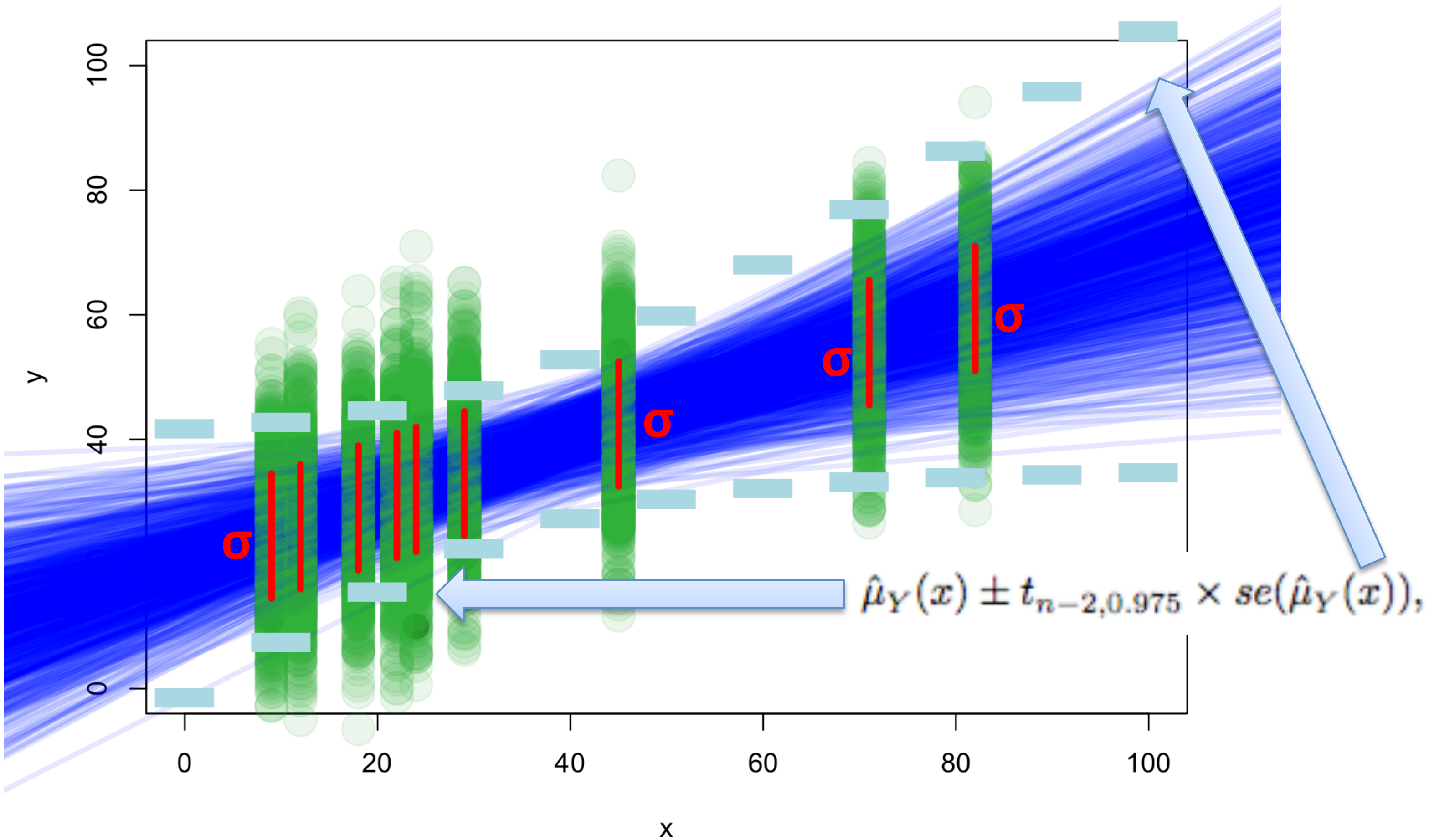
```
> # plot the variance for our different values of x:
> for(myx in x){
+ lines(c(myx,myx),c((beta0 + beta1*myx)-sqrt(sigma2),(beta0 + beta1*myx)
+sqrt(sigma2)),col="red",lwd=4)
+ }
```



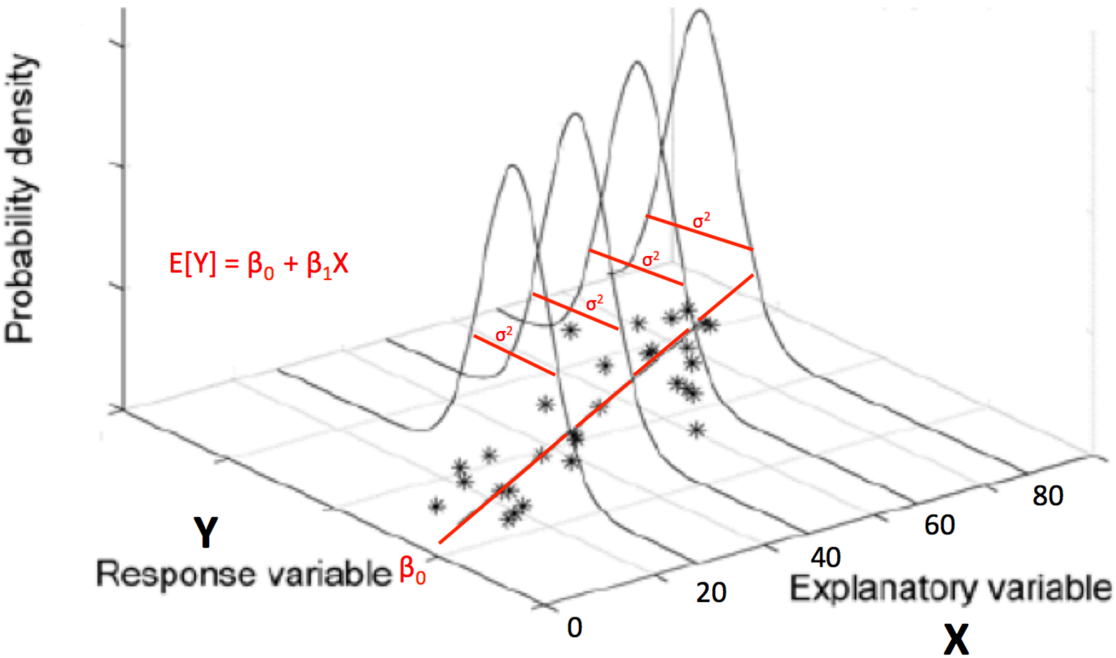
```
> # plot the variance for our different values of x:
> for(myx in x){
+ lines(c(myx,myx),c((beta0 + beta1*myx)-sqrt(sigma2),(beta0 + beta1*myx)
+sqrt(sigma2)),col="red",lwd=4)
+ }
```

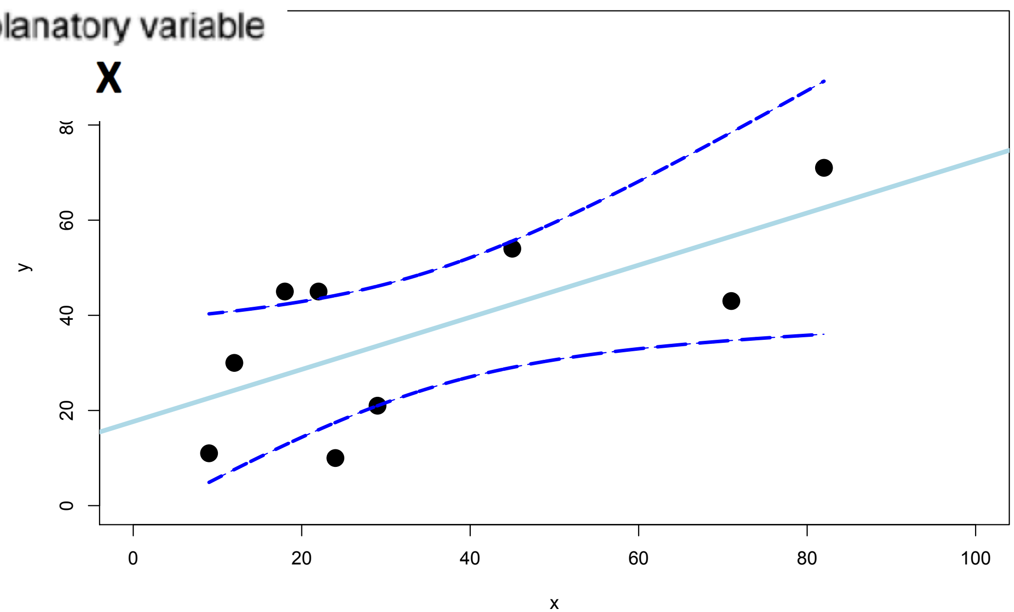


```
> # plot the variance for our different values of x:
> for(myx in x){
+ lines(c(myx,myx),c((beta0 + beta1*myx)-sqrt(sigma2),(beta0 + beta1*myx)
+sqrt(sigma2)),col="red",lwd=4)
+ }
```



- Confused about homogeneity vs. non-consistent width of confidence intervals?

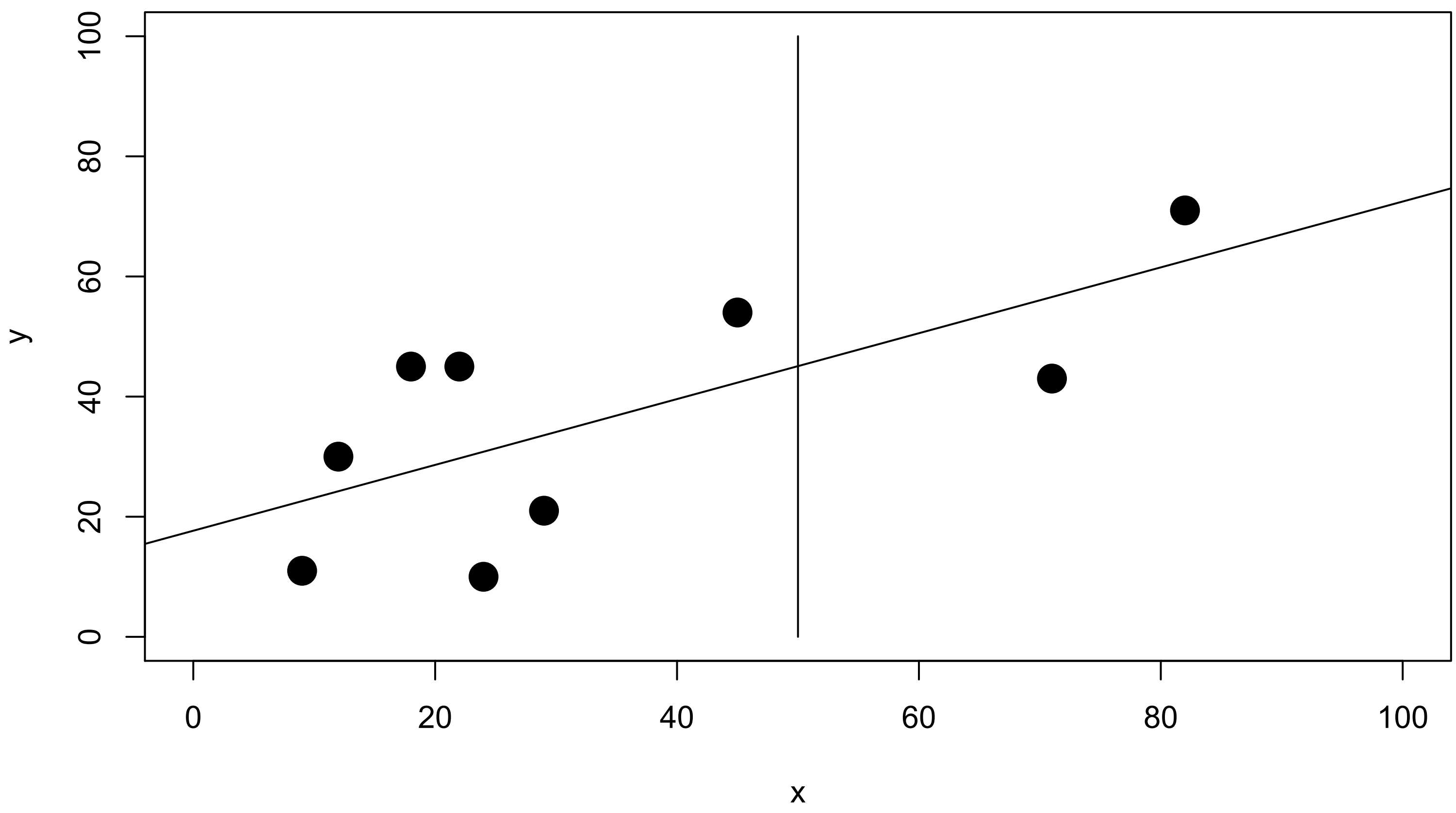




- Predictions and prediction intervals---

Suppose we now want to make a prediction for a new value of x.

Example: Suppose we would like to predict how much money (Y), someone aged 50 years old (X=50) will have.



- Predictions and prediction intervals---

Example: Suppose we would like to predict how much money (Y), someone aged X=50 years old will have.

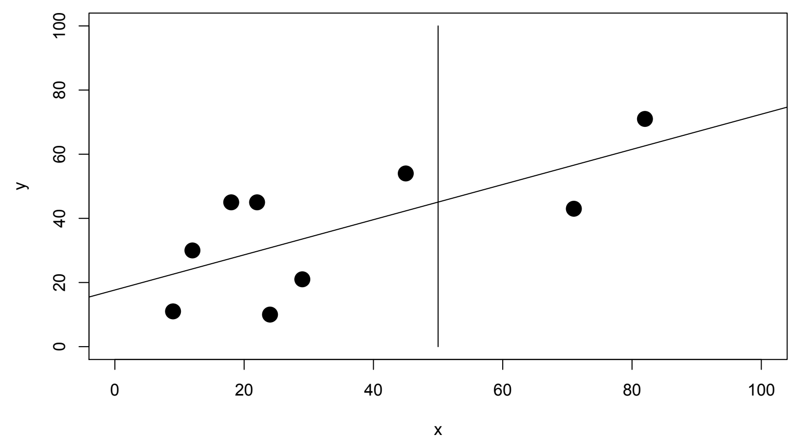
this hypothetical new person aged 50 is sometimes called “an out-of-sample unit with value x*”, Where x*=50.

Our best estimate, also known as the “point prediction”, would be equal to $b_0 + b_1(50) = 45.1$

```
> xstar <- 50
> point_prediction <- beta0hat + beta1hat*xstar
> point_prediction
[1] 45.07117
```

- Predictions and prediction intervals---

```
> # x and n are fixed values
> x <- c(82, 45, 71, 22, 29, 9, 12, 18, 24)
> n <- 9
>
> # y is a realization of the random variable "Y", i.e. "observed data":
> y <- c(71, 54, 43, 45, 21, 11, 30, 45, 10)
> xbar <- (1/n)*sum(x)
> ybar <- (1/n)*sum(y)
> sx <- sqrt( sum((x-xbar)^2)/(n-1) )
> sy <- sqrt( sum((y-ybar)^2)/(n-1) )
> sxy <- (1/(n-1))*sum((x-xbar)*(y-ybar))
> rxy <- sxy/(sx*sy)
> beta1hat <- rxy*sy/sx
> beta0hat <- ybar-beta1hat*xbar
> residuals <- y - beta0hat - beta1hat*x
> s <- sqrt( (1/(n-2))*sum(residuals^2))
> plot(y~x, xlim=c(0,100), ylim=c(0,100), pch=20, cex=3)
> abline(beta0hat, beta1hat)
>
> xstar <- 50
> point_prediction <- beta0hat + beta1hat*xstar
> point_prediction
[1] 45.07117
> lines(x=c(xstar, xstar),c(0,100))
```



- Predictions and prediction intervals---

Example: Suppose we would like to predict how much money (Y), someone aged X=60 years old will have.

$\hat{Y}(x^*) = \hat{B}_0 + \hat{B}_1 x^*$ with error

$$(2.67) \quad \hat{Y}(x^*) - Y(x^*) = \hat{B}_0 + \hat{B}_1 x^* - [\beta_0 + \beta_1 x^* + \epsilon(x^*)]$$

$$= (\hat{B}_0 - \beta_0) + (\hat{B}_1 - \beta_1) x^* - \epsilon(x^*)$$

- Predictions and prediction intervals---

Example: Suppose we would like to predict how much money (Y), someone aged X=60 years old will have.

Our prediction

The truth

$\hat{Y}(x^*) = \hat{B}_0 + \hat{B}_1 x^*$ with error

$$(2.67) \quad \hat{Y}(x^*) - Y(x^*) = \hat{B}_0 + \hat{B}_1 x^* - [\beta_0 + \beta_1 x^* + \epsilon(x^*)]$$

$$= (\hat{B}_0 - \beta_0) + (\hat{B}_1 - \beta_1)x^* - \epsilon(x^*)$$

The difference between our prediction and the truth is the error

- Predictions and prediction intervals---

Example: Suppose we would like to predict how much money (Y), someone aged X=60 years old will have.

Our prediction

The truth

$\hat{Y}(x^*) = \hat{B}_0 + \hat{B}_1 x^*$ with error

$$(2.67) \quad \hat{Y}(x^*) - Y(x^*) = \hat{B}_0 + \hat{B}_1 x^* - [\beta_0 + \beta_1 x^* + \epsilon(x^*)]$$

$$= (\hat{B}_0 - \beta_0) + (\hat{B}_1 - \beta_1)x^* - \epsilon(x^*)$$

The difference between our prediction and the truth is the error

This has variance

$$(2.68) \qquad \mathrm{Var}\left[(\hat{B}_0 - \beta_0) + (\hat{B}_1 - \beta_1)x^*\right] + \mathrm{Var}\left[\epsilon(x^*)\right] = \sigma^2\left\{n^{-1} + \frac{(x^* - \overline{x})^2}{[(n-1)s_x^2]}\right\} + \sigma^2,$$

- Predictions and prediction intervals---

Example: Suppose we would like to predict how much money (Y), someone aged X=60 years old will have.

Our prediction

The truth

$\hat{Y}(x^*) = \hat{B}_0 + \hat{B}_1 x^*$ with error

$$(2.67) \quad \hat{Y}(x^*) - Y(x^*) = \hat{B}_0 + \hat{B}_1 x^* - [\beta_0 + \beta_1 x^* + \epsilon(x^*)]$$
$$= (\hat{B}_0 - \beta_0) + (\hat{B}_1 - \beta_1)x^* - \epsilon(x^*)$$

The difference between our prediction and the truth is the error

Cov() is equal to 0, since the two terms are independent.

This has variance

$$(2.68) \quad \mathrm{Var}\left[(\hat{B}_0 - \beta_0) + (\hat{B}_1 - \beta_1)x^*\right] + \mathrm{Var}\left[\epsilon(x^*)\right] = \sigma^2\left\{n^{-1} + \frac{(x^* - \overline{x})^2}{[(n-1)s_x^2]}\right\} + \sigma^2,$$

since $\mathrm{Var}\left[(\hat{B}_0 - \beta_0) + (\hat{B}_1 - \beta_1)x^*\right] = \mathrm{Var}\left[\hat{\mu}_Y(x^*)\right]$ from (2.66).

- # Predictions and prediction intervals---

Example: Suppose we would like to predict how much money (Y), someone aged X=60 years old will have.

Our prediction

The truth

$\hat{Y}(x^*) = \hat{B}_0 + \hat{B}_1 x^*$ with error

$$(2.67) \quad \hat{Y}(x^*) - Y(x^*) = \hat{B}_0 + \hat{B}_1 x^* - [\beta_0 + \beta_1 x^* + \epsilon(x^*)]$$

$$= (\hat{B}_0 - \beta_0) + (\hat{B}_1 - \beta_1)x^* - \epsilon(x^*)$$

This has variance

$$(2.68) \quad \mathrm{Var}\,[(\hat{B}_0 - \beta_0) + (\hat{B}_1 - \beta_1)x^*] + \mathrm{Var}\,[\epsilon(x^*)] = \sigma^2\left\{n^{-1} + \frac{(x^* - \overline{x})^2}{[(n-1)s_x^2]}\right\} + \sigma^2,$$

So the (estimated) SE of the prediction error is

$$\hat{\sigma} \times \sqrt{1 + \frac{1}{n} + \frac{(x^* - \overline{x})^2}{(n-1)s_x^2}},$$

Note this does not decrease to 0 as $n \to \infty$.

- Predictions and prediction intervals---

Example: Suppose we would like to predict how much money (Y), someone aged X=60 years old will have.

Our prediction

The truth

$\hat{Y}(x^*) = \hat{B}_0 + \hat{B}_1 x^*$ with error

$$(2.67) \quad \hat{Y}(x^*) - Y(x^*) = \hat{B}_0 + \hat{B}_1 x^* - [\beta_0 + \beta_1 x^* + \epsilon(x^*)]$$

$$= (\hat{B}_0 - \beta_0) + (\hat{B}_1 - \beta_1)x^* - \epsilon(x^*)$$

This has variance

$$(2.68) \qquad \mathrm{Var}\,[(\hat{B}_0 - \beta_0) + (\hat{B}_1 - \beta_1)x^*] + \mathrm{Var}\,[\epsilon(x^*)] = \sigma^2\left\{n^{-1} + \frac{(x^* - \overline{x})^2}{[(n-1)s_x^2]}\right\} + \sigma^2,$$

So the (estimated) SE of the prediction error is

$$\hat{\sigma} \times \sqrt{1 + \frac{1}{n} + \frac{(x^* - \overline{x})^2}{(n-1)s_x^2}},$$

Note this does not decrease to 0 as $n \to \infty$.

Note that variances of estimators include σ^2 in their equations. Estimated SEs replace the "population" quantity σ by a sample quantity $\hat{\sigma}$.

- Predictions and prediction intervals---

Next for the 95% prediction interval for $Y(x^*)$ for an out-of-sample unit with value x^*, the point prediction is $\hat{Y}(x^*) = \hat{B}_0 + \hat{B}_1 x^*$ with error

$$\text{(2.67)} \quad \hat{Y}(x^*) - Y(x^*) = \hat{B}_0 + \hat{B}_1 x^* - [\beta_0 + \beta_1 x^* + \epsilon(x^*)] = (\hat{B}_0 - \beta_0) + (\hat{B}_1 - \beta_1)x^* - \epsilon(x^*).$$

This has variance

$$\text{(2.68)} \quad \mathrm{Var}\,[(\hat{B}_0 - \beta_0) + (\hat{B}_1 - \beta_1)x^*] + \mathrm{Var}\,[\epsilon(x^*)] = \sigma^2\left\{n^{-1} + \frac{(x^* - \overline{x})^2}{[(n-1)s_x^2]}\right\} + \sigma^2,$$

since $\mathrm{Var}\,[(\hat{B}_0 - \beta_0) + (\hat{B}_1 - \beta_1)x^*] = \mathrm{Var}\,[\hat{\mu}_Y(x^*)]$ from (2.66). So the (estimated) SE of the prediction error is

$$\text{(2.69)} \quad \hat{\sigma} \times \sqrt{1 + \frac{1}{n} + \frac{(x^* - \overline{x})^2}{(n-1)s_x^2}},$$

and this does not decrease to 0 as $n \to \infty$.

Note that variances of estimators include σ^2 in their equations. Estimated SEs replace the "population" quantity σ by a sample quantity $\hat{\sigma}$.

- Predictions and prediction intervals---

Next for the 95% prediction interval for $Y(x^*)$ for an out-of-sample unit with value x^*, the point prediction is $\hat{Y}(x^*) = \hat{B}_0 + \hat{B}_1 x^*$ with error

$$(2.67) \quad \hat{Y}(x^*) - Y(x^*) = \hat{B}_0 + \hat{B}_1 x^* - [\beta_0 + \beta_1 x^* + \epsilon(x^*)] = (\hat{B}_0 - \beta_0) + (\hat{B}_1 - \beta_1)x^* - \epsilon(x^*).$$

This has variance

$$(2.68) \qquad \mathrm{Var}\,[(\hat{B}_0 - \beta_0) + (\hat{B}_1 - \beta_1)x^*] + \mathrm{Var}\,[\epsilon(x^*)] = \sigma^2 \left\{ n^{-1} + \frac{(x^* - \overline{x})^2}{[(n-1)s_x^2]} \right\} + \sigma^2,$$

since $\mathrm{Var}\,[(\hat{B}_0 - \beta_0) + (\hat{B}_1 - \beta_1)x^*] = \mathrm{Var}\,[\hat{\mu}_Y(x^*)]$ from (2.66). So the (estimated) SE of the prediction error is

$$(2.69) \qquad \text{se(E)} = \quad \hat{\sigma} \times \sqrt{1 + \frac{1}{n} + \frac{(x^* - \overline{x})^2}{(n-1)s_x^2}},$$

and this does not decrease to 0 as $n \to \infty$.

Note that variances of estimators include σ^2 in their equations. Estimated SEs replace the "population" quantity σ by a sample quantity $\hat{\sigma}$.

- Predictions and prediction intervals---

Next for the 95% prediction interval for $Y(x^*)$ for an out-of-sample unit with value x^*, the point prediction is $\hat{Y}(x^*) = \hat{B}_0 + \hat{B}_1 x^*$ with error

(2.67) $\hat{Y}(x^*) - Y(x^*) = \hat{B}_0 + \hat{B}_1 x^* - [\beta_0 + \beta_1 x^* + \epsilon(x^*)] = (\hat{B}_0 - \beta_0) + (\hat{B}_1 - \beta_1)x^* - \epsilon(x^*).$

This has variance

(2.68) $$\mathrm{Var}\,[(\hat{B}_0 - \beta_0) + (\hat{B}_1 - \beta_1)x^*] + \mathrm{Var}\,[\epsilon(x^*)] = \sigma^2 \Big\{ n^{-1} + \frac{(x^* - \overline{x})^2}{[(n-1)s_x^2]} \Big\} + \sigma^2,$$

since $\mathrm{Var}\,[(\hat{B}_0 - \beta_0) + (\hat{B}_1 - \beta_1)x^*] = \mathrm{Var}\,[\hat{\mu}_Y(x^*)]$ from (2.66). So the (estimated) SE of the prediction error is

(2.69) $$\mathrm{se(E)} = \quad \hat{\sigma} \times \sqrt{1 + \frac{1}{n} + \frac{(x^* - \overline{x})^2}{(n-1)s_x^2}},$$

and this does not decrease to 0 as $n \to \infty$.

Note that variances of estimators include σ^2 in their equations. Estimated SEs replace the "population" quantity σ by a sample quantity $\hat{\sigma}$.

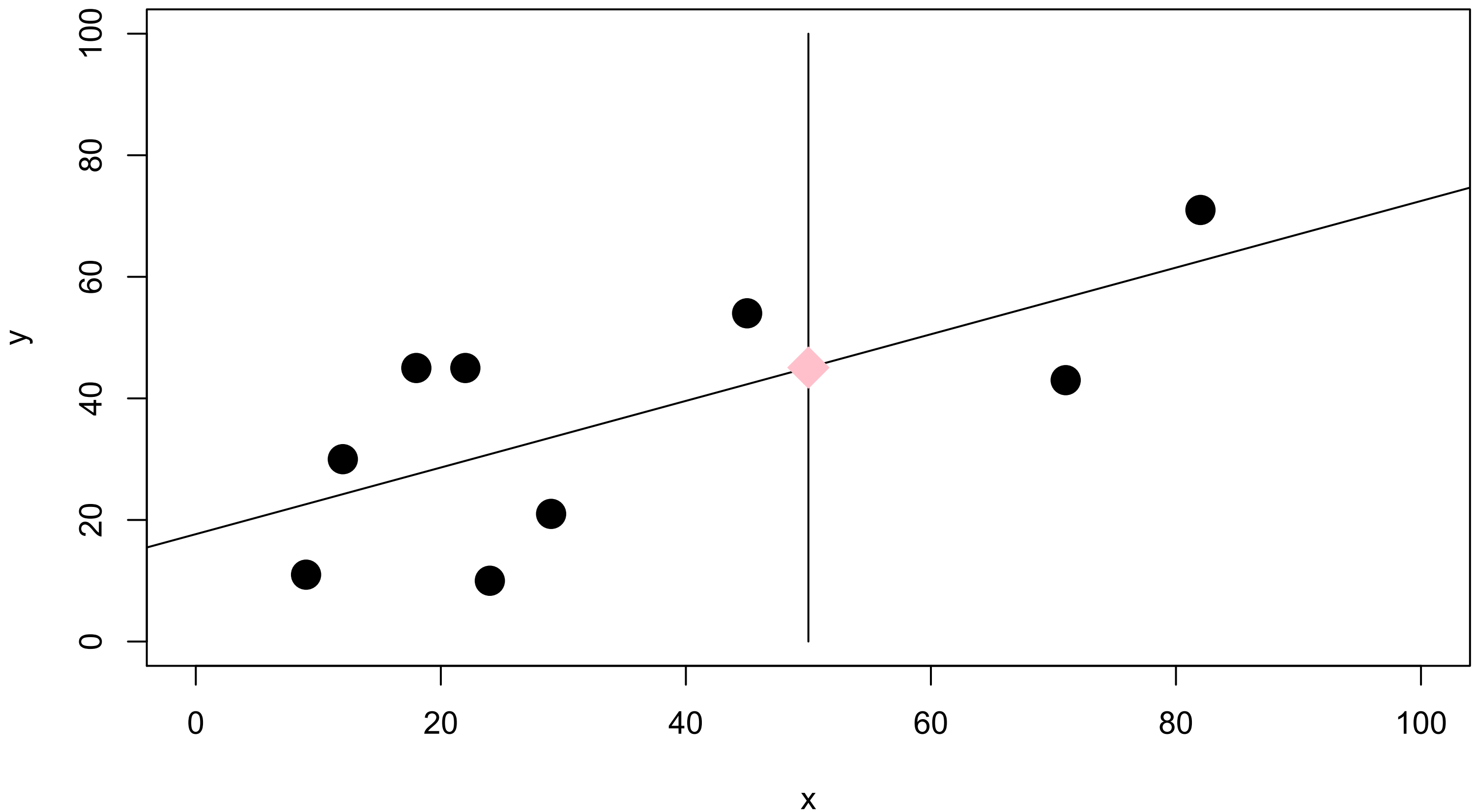
The 95% prediction interval for $Y(x^*)$ for a unit (not in sample) with value x^*:

(2.44) $$\hat{Y}(x^*) \pm t_{n-2,0.975} \times se(E),\ \ \hat{Y}(x^*) = \hat{\beta}_0 + \hat{\beta}_1 x^* = \hat{\mu}_Y(x^*),$$

where $E = \hat{Y}(x^*) - Y(x^*) = \hat{\mu}_Y(x^*) - Y(x^*) = \hat{\mu}_Y(x^*) - \beta_0 - \beta_1 x^* - \epsilon(x^*)$ is the prediction error.

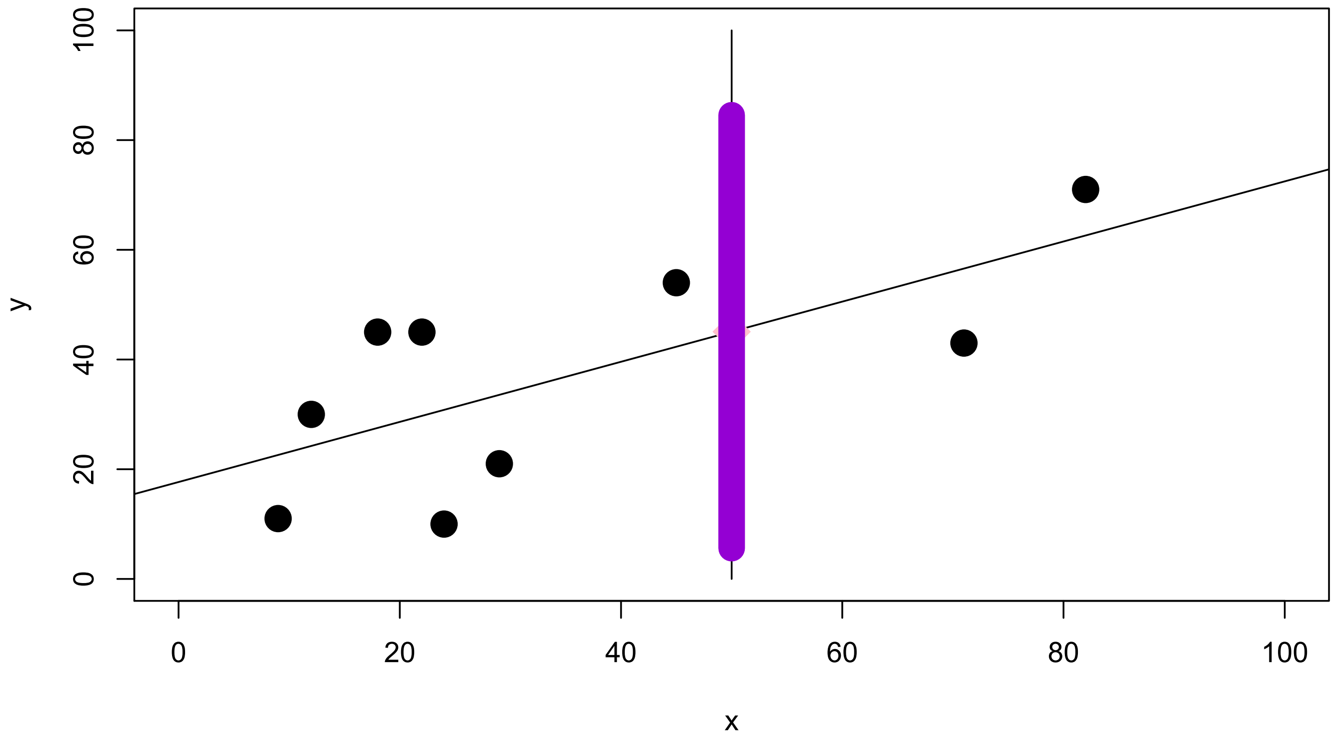
- Predictions and prediction intervals---

```
> points(xstar, point_prediction, col="pink", pch=18, cex=3)
```



- Predictions and prediction intervals---

```
> # 95% prediction interval:
> lowerPI <- point_prediction -  qt(0.975,n-2) * s * sqrt(1/n + 1 + ((xstar-xbar)^2)/((n-1)*sx^2))
> upperPI <- point_prediction +  qt(0.975,n-2) * s * sqrt(1/n + 1 + ((xstar-xbar)^2)/((n-1)*sx^2))
>
> c(lowerPI,upperPI)
[1]  5.61226 84.53007
>
> lines(x=c(xstar,xstar),y=c(lowerPI,upperPI), col="darkviolet",lwd=15)
```



Age vs. Money

Sample statistics

b_0 = 17.7
b_1 = 0.55
s = 15.5
R^2 = 0.49

Objective: The purpose of this observational study was to demonstrate if, and to what extent, age is associated with money.

Design and Methods: We collected a random sample of individuals and for each determined their age **(recorded in years)** and the amount of money (in dollars) in their accounts. Analysis of the data was done using **linear regression**.

For parameter β_1 :

$$95\% \text{ C.I.} = [0.05,\ 1.05]$$
$$p\text{-value} = 0.036$$

Results: We obtained a random sample of n = 9 subjects. There is a statistically significant association between age and money (p-value =0.036). For every additional year in age, an individual's amount of money increases on average by an estimated of $0.55 (95% C.I. = [$0.05, $1.05]).

Conclusions: We found that, as hypothesized, age is associated with money. In our sample age accounted for about half of the variability observed in money (R^2=0.49). **We predict that a 50 year old will have $45.1 (95% P.I. = [$5.6, $84.5]),** whereas a 40 year old will have $39.6 (95% P.I. = [$0.8, $78.4]).

Small Print: The analysis rests on the following assumptions:

- the observations are independently and identically distributed.
- the **response** variable, money, is normally distributed.
- Homoscedasticity of residuals or equal variance.
- the relationship between **response** and **predictor** variables is linear.

Age vs. Money

Sample statistics

b_0 = 17.7
b_1 = 0.55
s = 15.5
R^2 = 0.49

For parameter β_1 :

$$95\% \text{ C.I.} = [0.05, 1.05]$$
$$p\text{-value} = 0.036$$

Objective: The purpose of this observational study was to demonstrate if, and to what extent, age is associated with money.

Design and Methods: We collected a random sample of individuals and for each determined their age **(recorded in years)** and the amount of money (in dollars) in their accounts. Analysis of the data was done using **linear regression**.

Results: We obtained a random sample of n = 9 subjects. There is a statistically significant association between age and money (p-value =0.036). For every additional year in age, an individual's amount of money increases on average by an estimated of \$0.55 (95% C.I. = [\$0.05, \$1.05]).

Conclusions: We found that, as hypothesized, age is associated with money. In our sample age accounted for about half of the variability observed in money (R^2=0.49). **We predict that a 50 year old will have \$45.1 (95% P.I. = [\$5.6, \$84.5]),** whereas a 40 year old will have \$39.6 (95% P.I. = [\$0.8, \$78.4]).

Small Print: The analysis rests on the following assumptions:

- the observations are independently and identically distributed.
- the **response** variable, money, is normally distributed.
- Homoscedasticity of residuals or equal variance.
- the relationship between **response** and **predictor** variables is linear.

Age vs. Money

Sample statistics

b_0 = 17.7
b_1 = 0.55
s = 15.5
R^2 = 0.49

For parameter β_1 :

$$95\% \text{ C.I.} = [0.05, 1.05]$$
$$p\text{-value} = 0.036$$

Objective: The purpose of this observational study was to demonstrate if, and to what extent, age is associated with money.

Design and Methods: We collected a random sample of individuals and for each determined their age **(recorded in years)** and the amount of money (in dollars) in their accounts. Analysis of the data was done using **<u>linear regression</u>**.

Results: We obtained a random sample of *n* = 9 subjects. There is a statistically significant association between age and money (*p*-value =0.036). For every additional year in age, an individual's amount of money increases on average by an estimated of $0.55 (95% C.I. = [$0.05, $1.05]).

Conclusions: We found that, as hypothesized, age is associated with money. In our sample age accounted for about half of the variability observed in money (R^2=0.49). **We <u>predict</u> that a 50 year old will have $45.1 (95% P.I. = [$5.6, $84.5]),** whereas a 40 year old will have $39.6 (95% P.I. = [$0.8, $78.4]).

Small Print: The analysis rests on the following assumptions:

- the observations are independently and identically distributed.
- the **response** variable, money, is normally distributed.
- Homoscedasticity of residuals or equal variance.
- the <u>relationship</u> between **response** and **predictor** variables is linear.

Age vs. Money

Sample statistics

$b_0 = 17.7$
$b_1 = 0.55$
$s = 15.5$
$R^2 = 0.49$

For parameter β_1 :

$$95\% \text{ C.I.} = [0.05, 1.05]$$
$$p\text{-value} = 0.036$$

Objective: The purpose of this observational study was to demonstrate if, and to what extent, age is associated with money.

Design and Methods: We collected a random sample of individuals and for each determined their age **(recorded in years)** and the amount of money (in dollars) in their accounts. Analysis of the data was done using **linear regression**.

Results: We obtained a random sample of $n = 9$ subjects. There is a statistically significant association between age and money (p-value =0.036). For every additional year in age, an individual's amount of money increases on average by an estimated of \$0.55 (95% C.I. = [\$0.05, \$1.05]).

Conclusions: We found that, as hypothesized, age is associated with money. In our sample age accounted for about half of the variability observed in money (R^2=0.49). **We predict that a 50 year old will have \$45.1 (95% P.I. = [\$5.6, \$84.5]),** whereas a 40 year old will have \$39.6 (95% P.I. = [\$0.8, \$78.4]).

Small Print: The analysis rests on the following assumptions:

- the observations are independently and identically distributed.
- the **response** variable, money, is normally distributed.
- Homoscedasticity of residuals or equal variance.
- the relationship between **response** and **predictor** variables is linear.

Age vs. Money

Sample statistics

b_0 = 17.7
b_1 = 0.55
s = 15.5
R^2 = 0.49

For parameter β_1 :

$$95\%\ \text{C.I.} = [0.05,\ 1.05]$$
$$p\text{-value} = 0.036$$

Objective: The purpose of this observational study was to demonstrate if, and to what extent, age is associated with money.

Design and Methods: We collected a random sample of individuals and for each determined their age **(recorded in years)** and the amount of money (in dollars) in their accounts. Analysis of the data was done using **linear regression**.

Results: We obtained a random sample of n = 9 subjects. There is a statistically significant association between age and money (p-value =0.036). For every additional year in age, an individual's amount of money increases on average by an estimated of \$0.55 (95% C.I. = [\$0.05, \$1.05]).

Conclusions: We found that, as hypothesized, age is associated with money. In our sample age accounted for about half of the variability observed in money (R^2=0.49). **We predict that a 50 year old will have \$45.1 (95% P.I. = [\$5.6, \$84.5]),** whereas a 40 year old will have \$39.6 (95% P.I. = [\$0.8, \$78.4]).

Small Print: The analysis rests on the following assumptions:

- the observations are independently and identically distributed.
- the **response** variable, money, is normally distributed.
- Homoscedasticity of residuals or equal variance.
- the relationship between **response** and **predictor** variables is linear.

- Questions?