Stat 306: Finding Relationships in Data.

Lecture 3

Residuals and 2.2 Statistical linear regression model

t-test

Age vs. Money

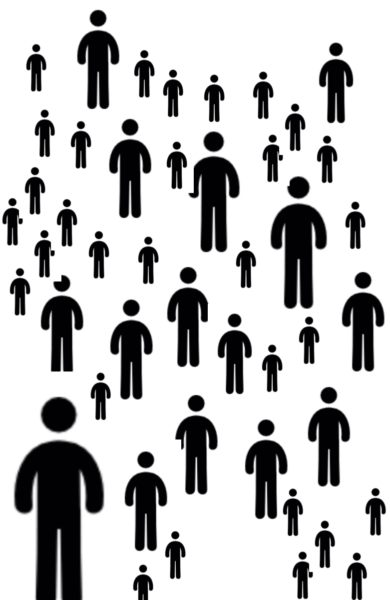
Dependent variable

X → old (0), young (1)

Independent variable

Y → dollars ($)
In bank account

Population



Population parameters

$$\mu_0, \mu_1, \sigma^2$$

Hypothesis Test

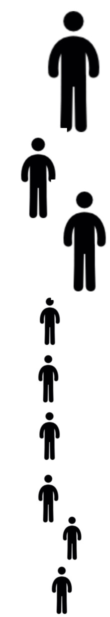
$$H_0 : \mu_0 = \mu_1$$
$$H_1 : \mu_0 \neq \mu_1$$

Sample statistics

$$\bar{y}_0 = 56$$
$$\bar{y}_1 = 27$$
$$\bar{y}_0 - \bar{y}_1 = 29$$
$$s_p = 10.81$$

t = 2.68, df = 7
p-value = 0.03
95% C.I. = [3.4, 54.6]

Sample, n=9



x	y
old	71
old	54
old	43
young	45
young	21
young	11
young	30
young	45
young	10

linear regression

Age vs. Money

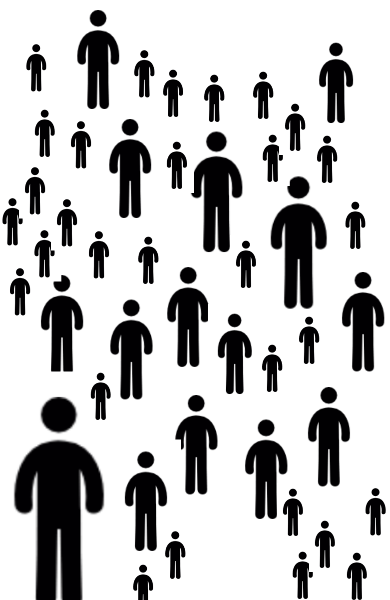
PREDICTOR variable

$X \longrightarrow$ Age in Years

RESPONSE variable

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

Population



Population parameters

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

Hypothesis Test

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

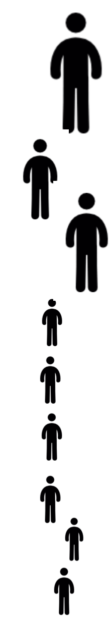
Sample statistics

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

For parameter β_1 :

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

Sample, n=9



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

Sample statistics

$$\bar{x} = n^{-1}\sum_{i=1}^{n} x_i, \quad s_x = \sqrt{\sum_{i=1}^{n}(x_i - \bar{x})^2 \,/\, (n-1)}\,,$$

$$\bar{y} = n^{-1}\sum_{i=1}^{n} y_i, \quad s_y = \sqrt{\sum_{i=1}^{n}(y_i - \bar{y})^2 \,/\, (n-1)}\,.$$

$$s_{xy} = (n-1)^{-1}\sum_{i=1}^{n}(x_i - \bar{x})(y_i - \bar{y}).$$

$$r_{xy} = \frac{s_{xy}}{s_x s_y}$$

Sample statistics

```
> x <- c(82, 45, 71, 22, 29, 9, 12, 18, 24)
> y <- c(71, 54, 43, 45, 21, 11, 30, 45, 10)
> n <- 9
```

Formulas as written in the course notes:

$$\bar{x} = n^{-1}\sum_{i=1}^{n} x_i, \quad s_x = \sqrt{\sum_{i=1}^{n}(x_i - \bar{x})^2 \,/\, (n-1)}\,,$$

$$\bar{y} = n^{-1}\sum_{i=1}^{n} y_i, \quad s_y = \sqrt{\sum_{i=1}^{n}(y_i - \bar{y})^2 \,/\, (n-1)}\,.$$

$$s_{xy} = (n-1)^{-1}\sum_{i=1}^{n}(x_i - \bar{x})(y_i - \bar{y}).$$

$$r_{xy} = \frac{s_{xy}}{s_x s_y}$$

Formulas written in R code:

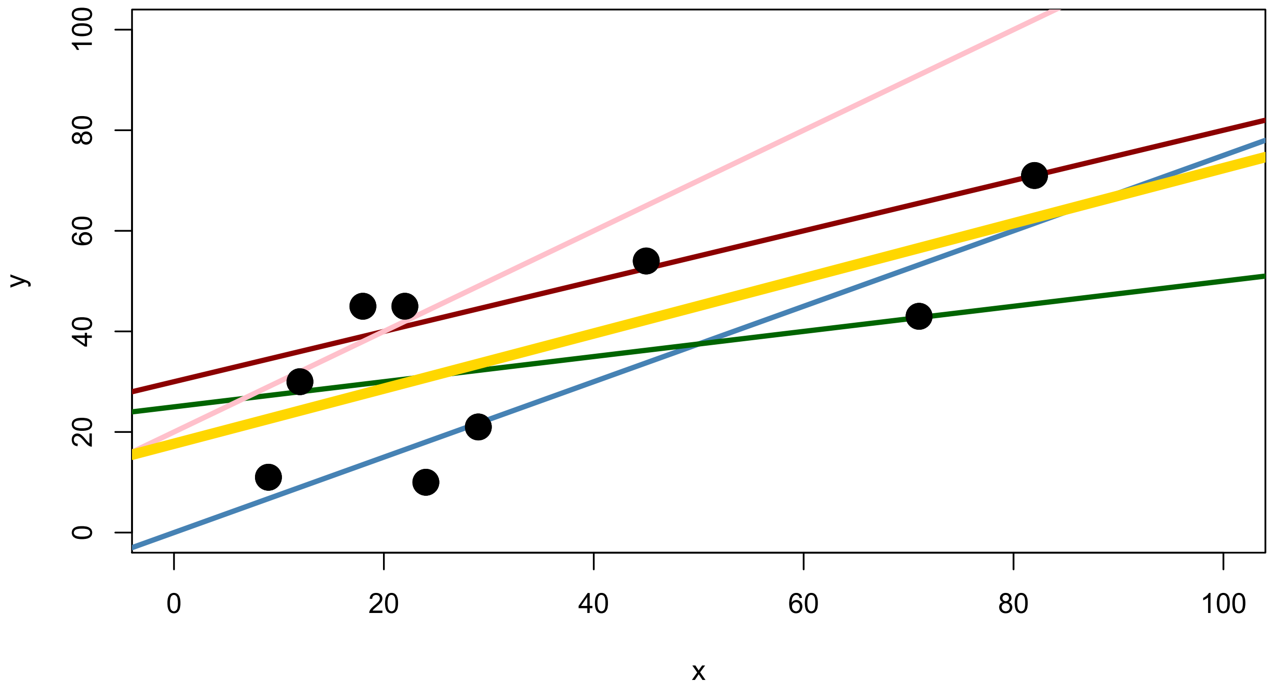
```
> xbar<-(1/n)*sum(x)
> xbar
[1] 34.66667
>
> sx<-sqrt( sum((x-xbar)^2)/(n-1) )
> sx
[1] 26.03843
>
> ybar<-(1/n)*sum(y)
> ybar
[1] 36.66667
>
> sy<-sqrt( sum((y-ybar)^2)/(n-1) )
> sy
[1] 20.36541
>
> sxy<-(1/(n-1))*sum((x-xbar)*(y-ybar))
> sxy
[1] 371.625
> rxy<-sxy/(sx*sy)
> rxy
[1] 0.7008045
```

The goal is to minimize $S(b_0, b_1) = \sum_{i=1}^{n}(y_i - b_0 - b_1 x_i)^2$.

Least Squares Solution:

$$\hat{b}_0 = \overline{y} - \hat{b}_1 \overline{x}$$

$$\hat{b}_1 = r_{xy} s_y / s_x$$



y = 0 + 1x   S(b0,b1) = 2933.5

y = 25 + 0.25x   S(b0,b1) = 2251.5

y = 30 + 0.5x   S(b0,b1) = 2725.0

y = 20 + 1x   S(b0,b1) = 5712.0

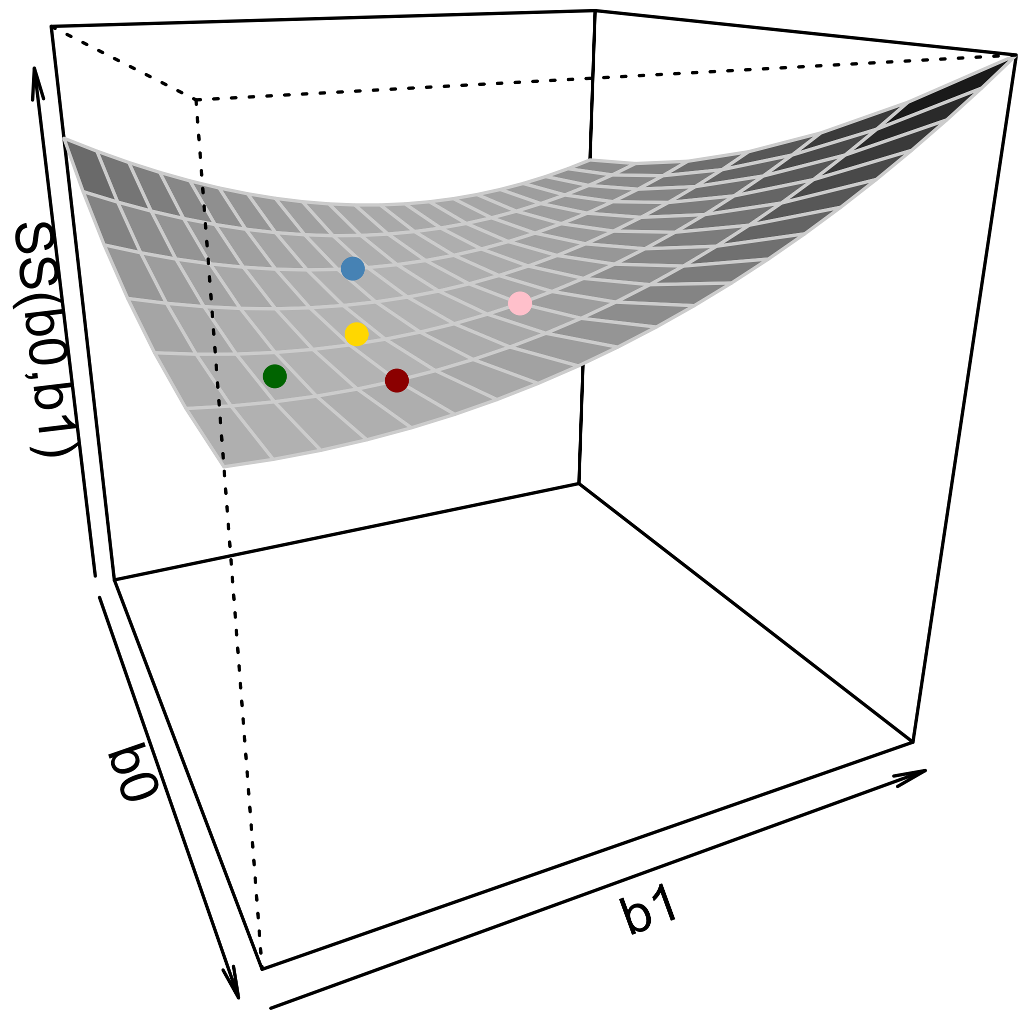
y = 17.7 + 0.55x   S(b0,b1) = 1688.4

The goal is to minimize $S(b_0, b_1) = \sum_{i=1}^{n}(y_i - b_0 - b_1 x_i)^2$.

Least Squares Solution:

$$\hat{b}_0 = \overline{y} - \hat{b}_1 \overline{x}$$

$$\hat{b}_1 = r_{xy} s_y / s_x$$



Least Squares Solution:

$$\hat{b}_1 = r_{xy} s_y / s_x$$

$$\hat{b}_0 = \overline{y} - \hat{b}_1 \overline{x}$$

```
> b1_hat<-rxy*sy/sx
> b0_hat<-ybar-b1_hat*xbar
>
> b1_hat
[1] 0.5481195
> b0_hat
[1] 17.66519
```

Least Squares Solution:

$$\hat{b}_1 = r_{xy} s_y / s_x$$

$$\hat{b}_0 = \overline{y} - \hat{b}_1 \overline{x}$$

```
> b1_hat<-rxy*sy/sx
> b0_hat<-ybar-b1_hat*xbar
>
> b1_hat
[1] 0.5481195
> b0_hat
[1] 17.66519
```

Predicted values:

$$\hat{y} = \hat{b}_0 + \hat{b}_1 x$$

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% C.I. $= [0.05,\ 1.05]$
p-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 <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 relationship between response and predictor variables is linear.

Least Squares Solution:

$$\hat{b}_0 = \overline{y} - \hat{b}_1 \overline{x}$$

$$\hat{b}_1 = r_{xy} s_y / s_x$$

Predicted values:

$$\hat{y} = \hat{b}_0 + \hat{b}_1 x$$

We predict that a 50 year old will have $45.1, whereas a 40 year old will have $39.6.

$$45.1 = 17.67 + 0.548 * 50$$

$$39.6 = 17.67 + 0.548 * 40$$

Least Squares Solution:

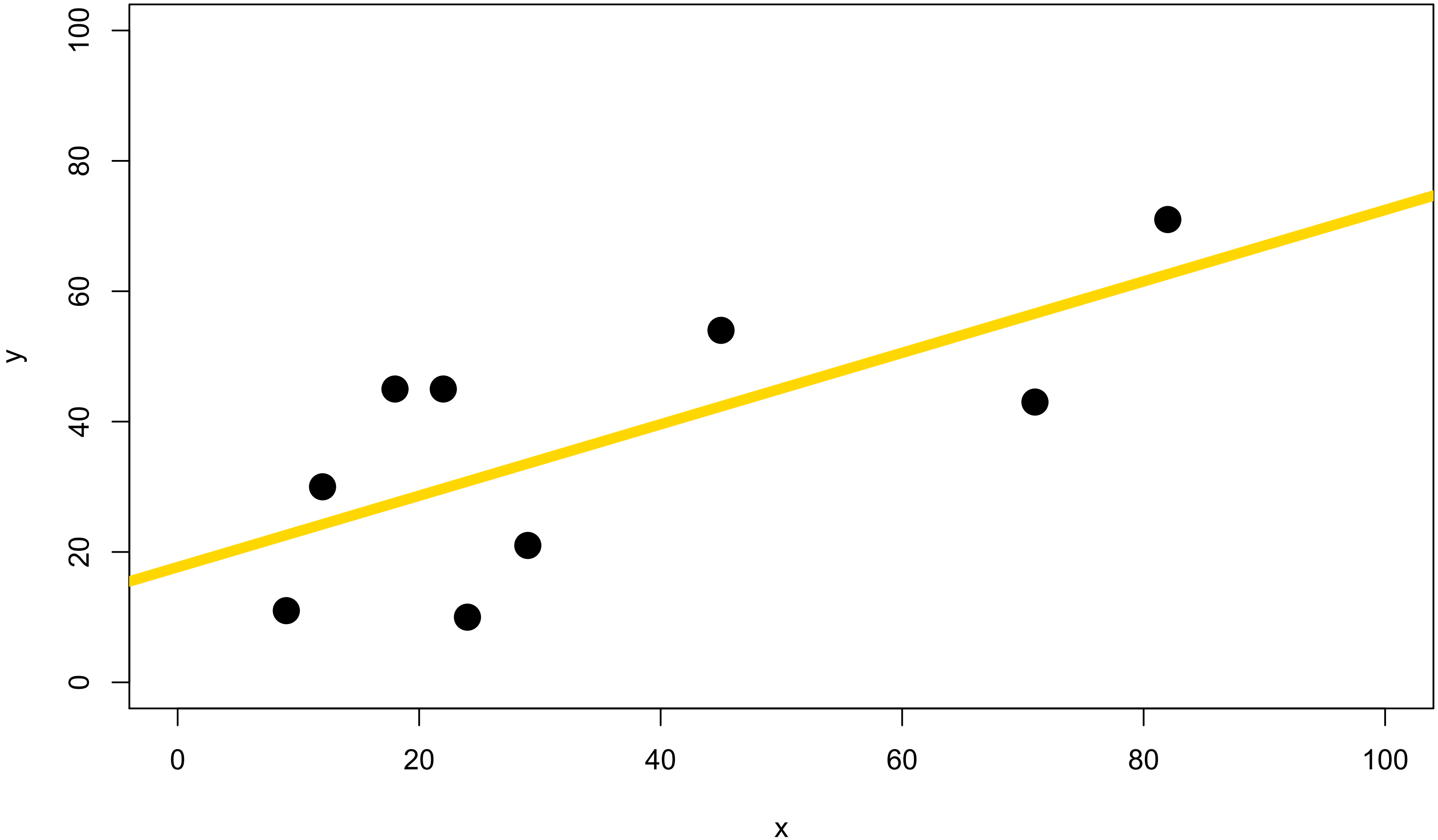
$$\hat{b}_0 = \overline{y} - \hat{b}_1\overline{x}$$

$$\hat{b}_1 = r_{xy}s_y/s_x$$

Predicted values:

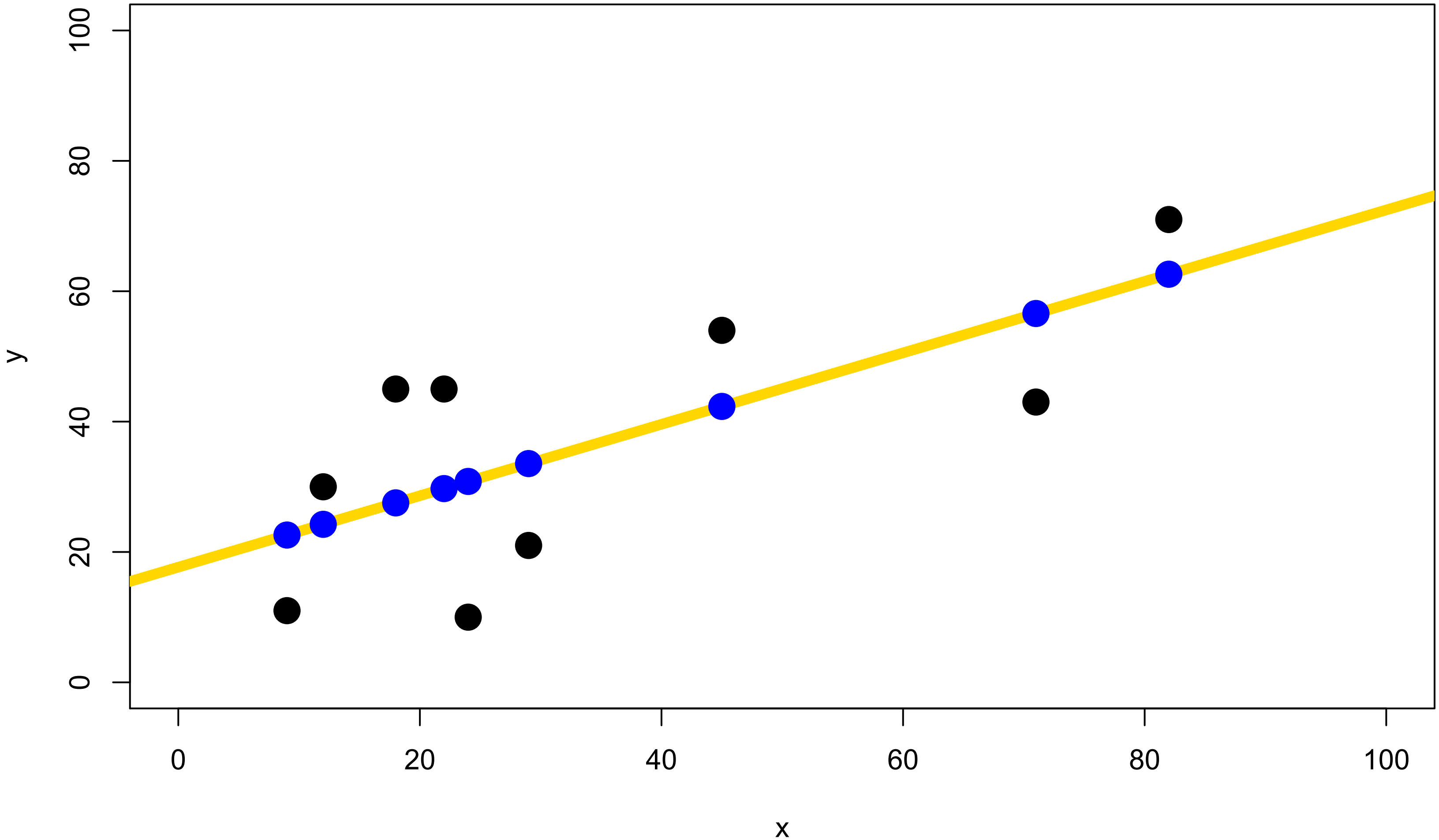
$$\hat{y} = \hat{b}_0 + \hat{b}_1 x$$

```
> yhat<-b0_hat+b1_hat*x
> yhat
[1] 62.61099 42.33057 56.58167 29.72382 33.56066 22.59827 24.24263
[8] 27.53134 30.82006
```




```
> plot(y~x, pch=20, cex=3, xlim=c(0,100), ylim=c(0,100))
> abline(17.67, 0.548 , col="gold", lwd=6)
```

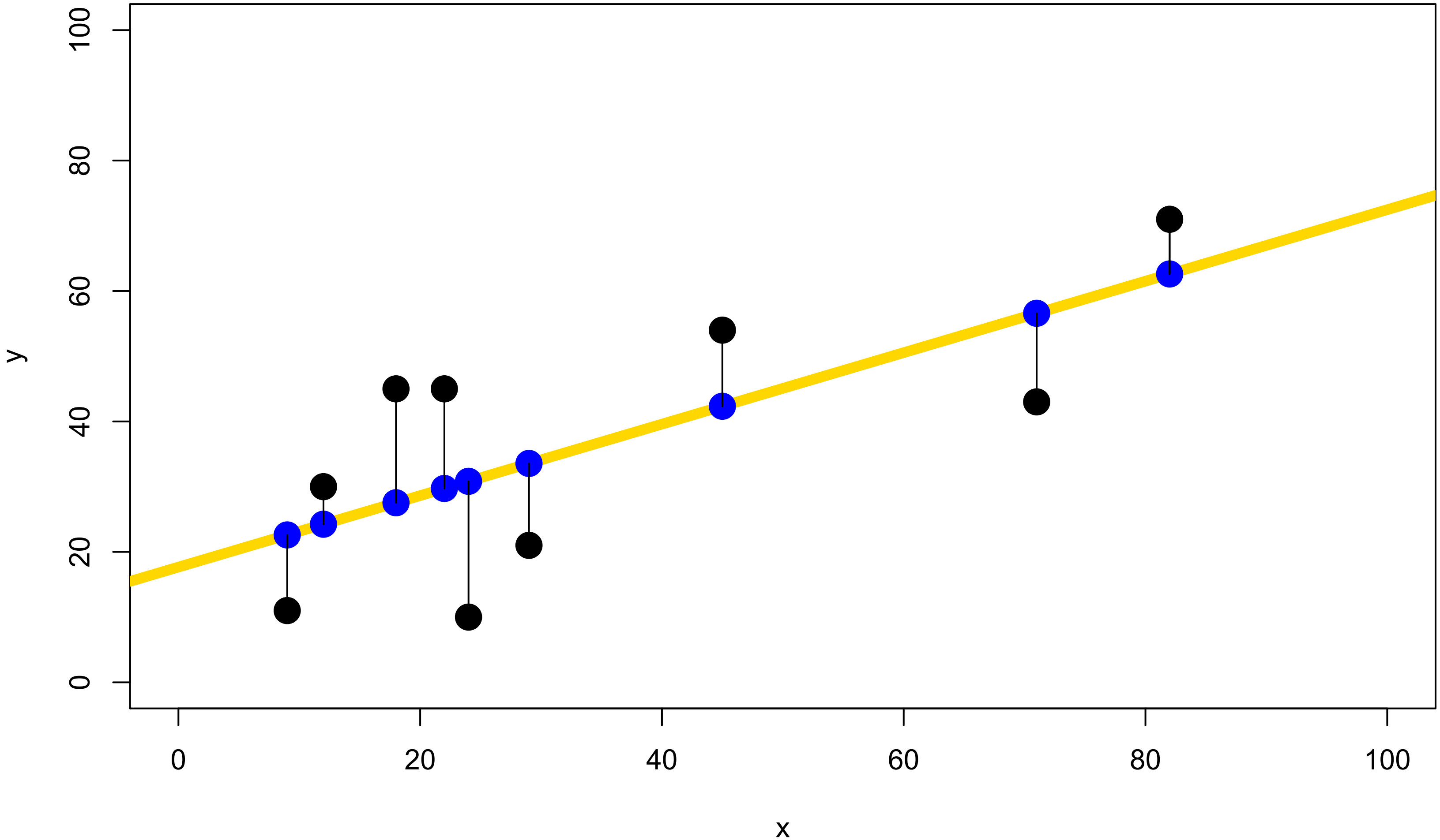
$$\hat{b}_0 = \overline{y} - \hat{b}_1 \overline{x} \qquad \hat{b}_1 = r_{xy} s_y / s_x$$



Predicted values:

$$\hat{y} = \hat{b}_0 + \hat{b}_1 x$$

```
> yhat<-b0_hat+b1_hat*x
> points(x,yhat, pch=20, cex=3, col="blue")
```



Residuals

$$e_i = y_i - \hat{b}_0 - \hat{b}_1 x_i, \quad i = 1, \ldots, n.$$

```
> residuals
[1]   8.389012  11.669432 -13.581674  15.276180 -12.560656 -11.598267
[7]   5.757375  17.468658 -20.820059
```

Residuals

$$e_i = y_i - \hat{b}_0 - \hat{b}_1 x_i, \quad i = 1, \ldots, n.$$

The goal is to minimize $S(b_0, b_1) = \sum_{i=1}^{n} (y_i - b_0 - b_1 x_i)^2$

We have: $e_i = y_i - \hat{b}_0 - \hat{b}_1 x_i, \quad i = 1, \ldots, n.$

Therefore:

$$\begin{aligned} S(b_0, b_1) &= \sum_{i=1}^{n} (\quad e_i \quad)^2 \\ &= \sum_{i=1}^{n} (y_i - b_0 - b_1 x_i)^2 \end{aligned}$$

3.8 Residual Plots

3.8 Residual plots

In this section, residual plots are introduced to check if the model (3.36) is an adequate approximation to (3.33), and also to check the normality and homoscedasticity assumptions.

the *residual*

$$e_i = y_i - \hat{b}_0 - \hat{b}_1 x_i, \quad i = 1, \ldots, n.$$

Residual plots include the following.

(i) Check for homoscedasticity versus heteroscedasticity and possible structural deviations from model (plot of residuals versus predicted values, plots of residuals versus each explanatory variable).

(ii) Check for normality (normal quantile plot of residuals) if the plots from (i) look OK.

3.8 Residual Plots

3.8 Residual plots

In this section, residual plots are introduced to check if the model (3.36) is an adequate approximation to (3.33), and also to check the normality and homoscedasticity assumptions.

the *residual*

$$e_i = y_i - \hat{b}_0 - \hat{b}_1 x_i, \quad i = 1, \ldots, n.$$

Residual plots include the following.

(i) Check for homoscedasticity versus heteroscedasticity and possible structural deviations from model (<u>plot of residuals versus predicted values</u>, <u>plots of residuals versus each explanatory variable</u>).

(1) (2)

(ii) Check for normality <u>(normal quantile plot of residuals)</u> if the plots from (i) look OK.

(3)

Age vs. Money

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

3.8 Residual Plots

(i) Check for homoscedasticity versus heteroscedasticity and possible structural deviations from model

Homoscedasticity: EQUAL VARIANCE

Heteroscedasticity: NOT EQUAL VARIANCE

3.8 Residual Plots

(i) Check for homoscedasticity versus heteroscedasticity and possible structural deviations from model

Homoscedasticity: EQUAL VARIANCE

Heteroscedasticity: NOT EQUAL VARIANCE

(i) Check for homoscedasticity versus heteroscedasticity and possible structural deviations from model (plot of residuals versus predicted values, plots of residuals versus each explanatory variable).

(1) (2)

(ii) Check for normality (normal quantile plot of residuals) if the plots from (i) look OK.

(3)

3.8 Residual Plots

Homoscedasticity of residuals or "equal variance"

(1) Plot of residuals versus predicted values.

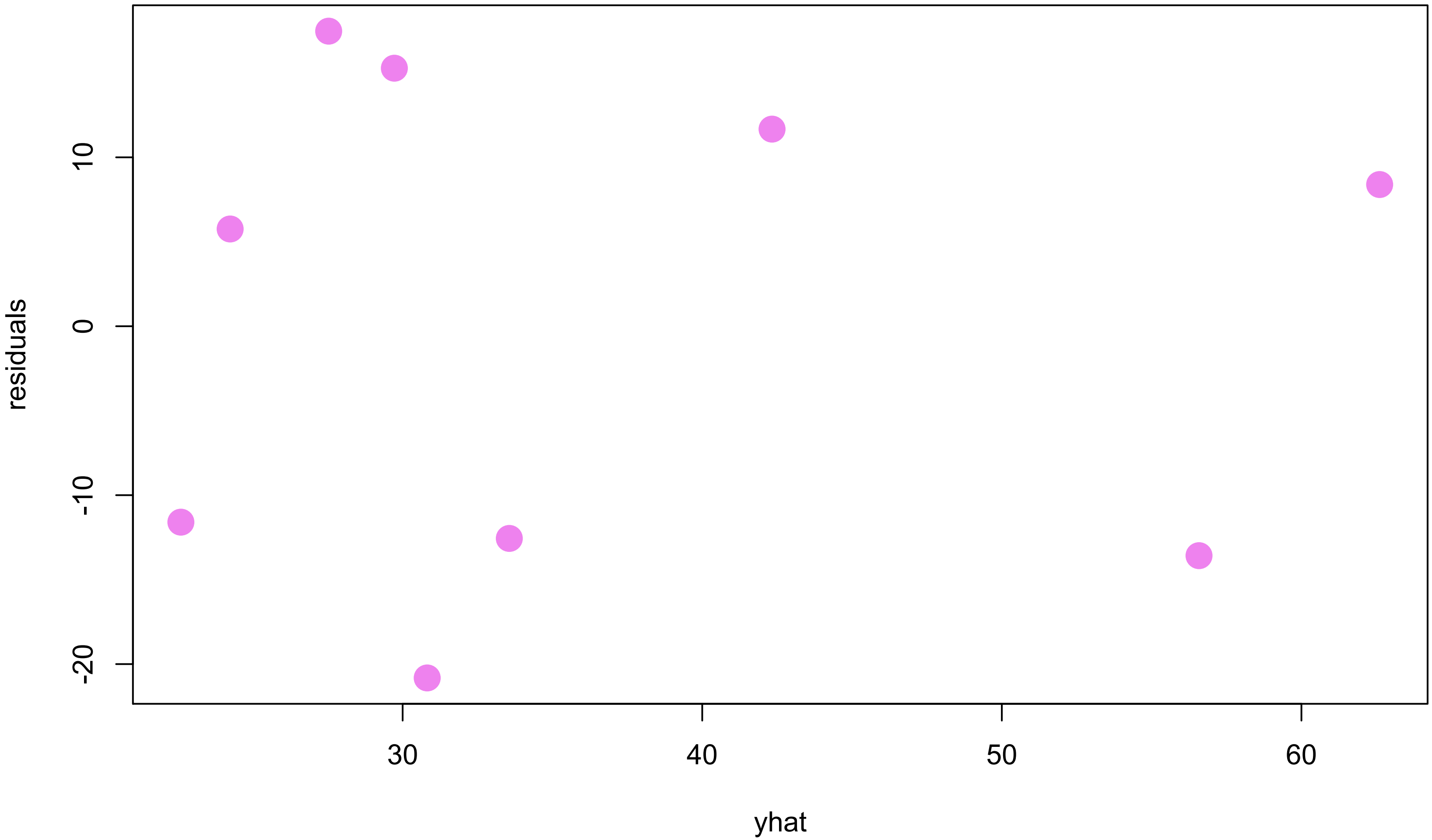
(2) Plot of residuals versus explanatory value

3.8 Residual Plots

Homoscedasticity of residuals or "equal variance"

(1) Plot of residuals versus predicted values:

```
plot(residuals~yhat, pch=20, cex=3, col="violet")
```

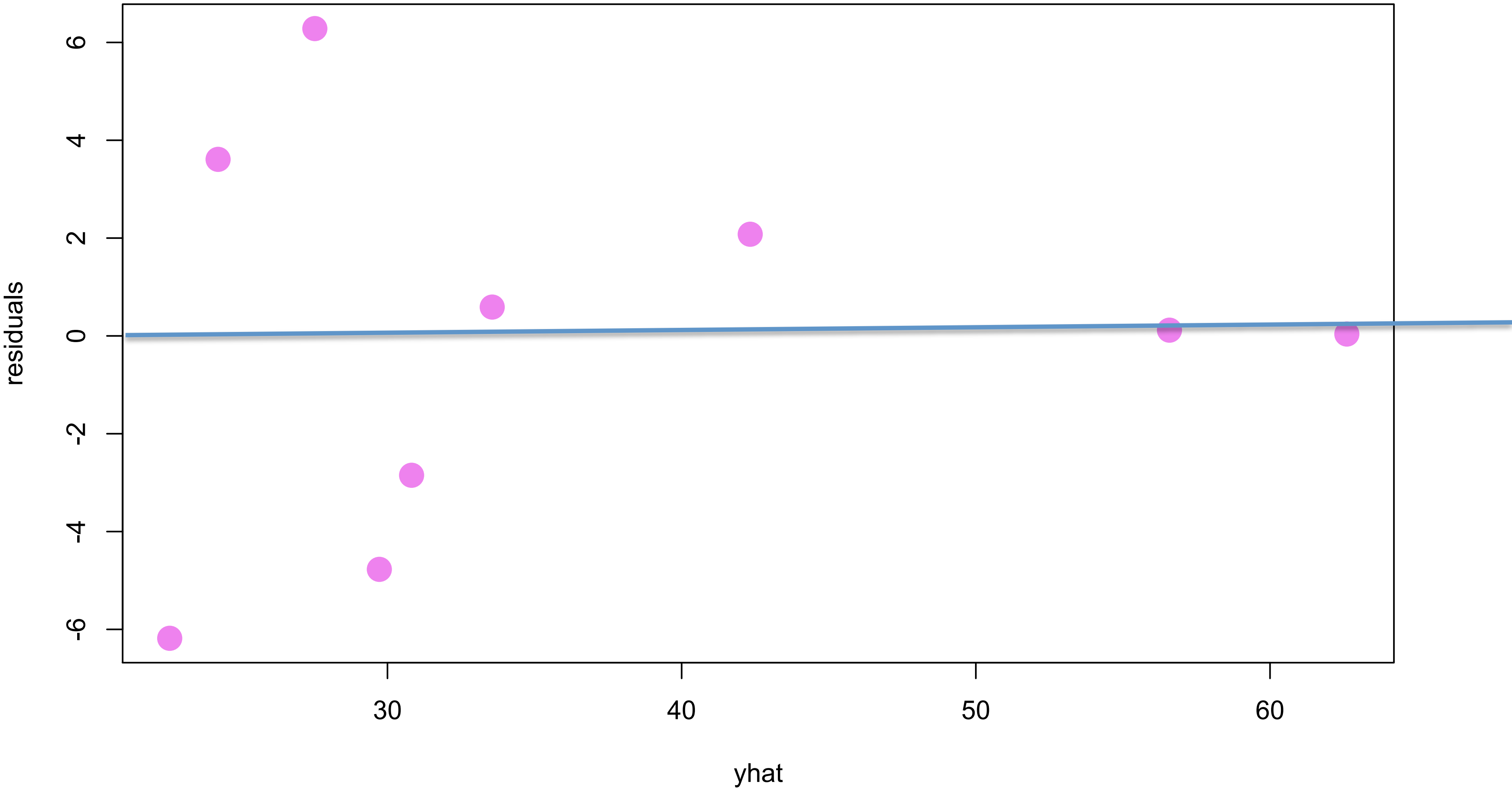


3.8 Residual Plots

Heteroscedasticity of residuals or "not equal variance"

(1) Plot of residuals versus predicted values:

```
plot(residuals~yhat, pch=20, cex=3, col="violet")
```

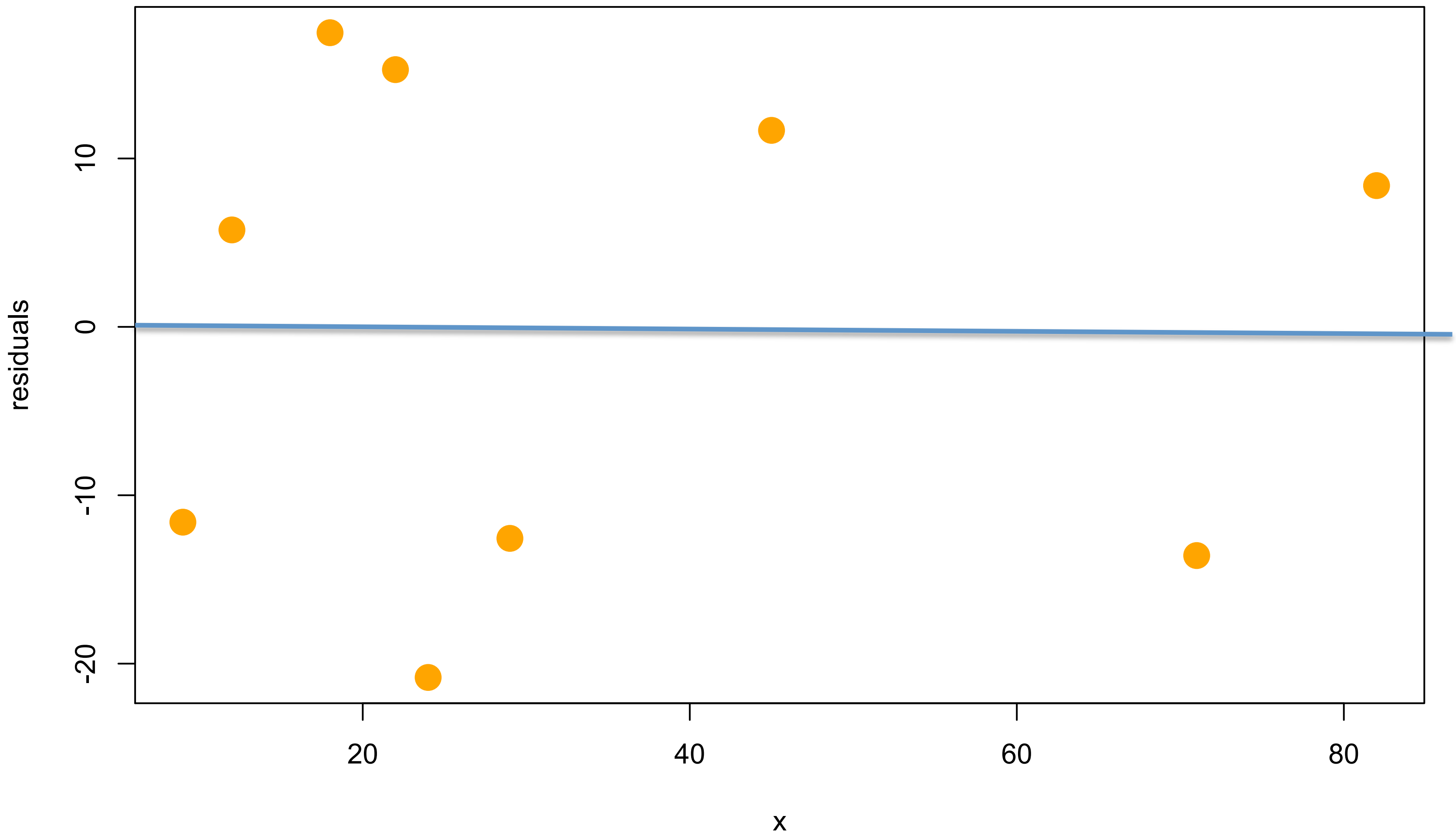


3.8 Residual Plots

Homoscedasticity of residuals or "equal variance"

(2) Plot of residuals versus explanatory value:

```
> plot(residuals~x, pch=20, cex=3, col="orange")
```

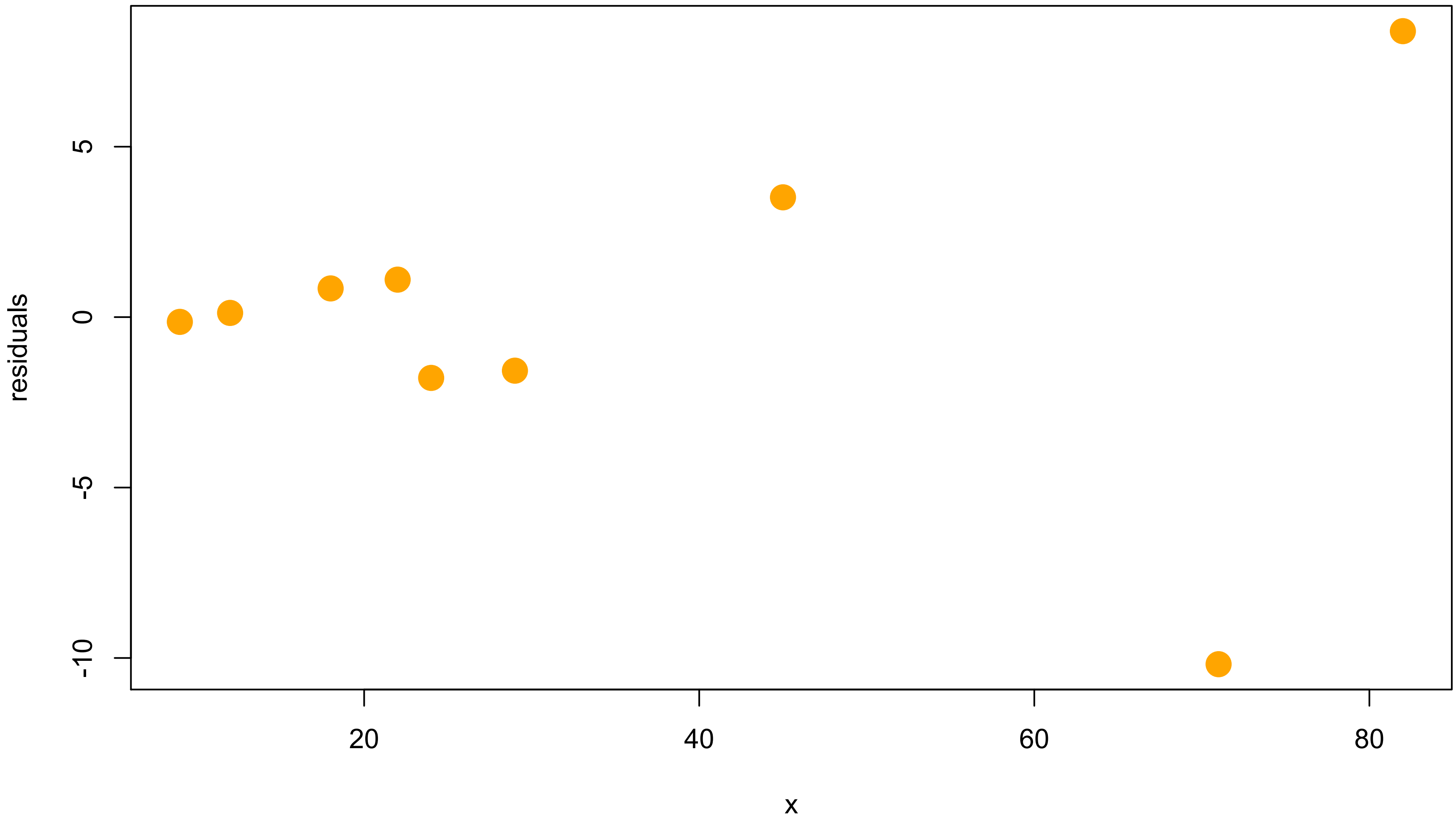


3.8 Residual Plots

Heteroscedasticity of residuals or "not equal variance"

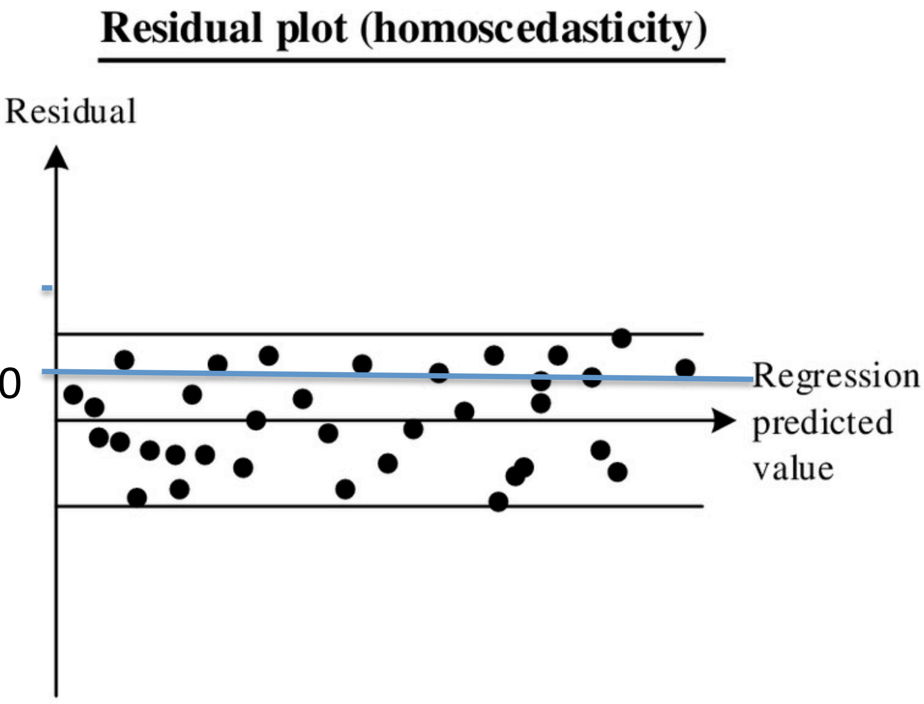
(2) Plot of residuals versus explanatory value:

```
> plot(residuals~x, pch=20, cex=3, col="orange")
```

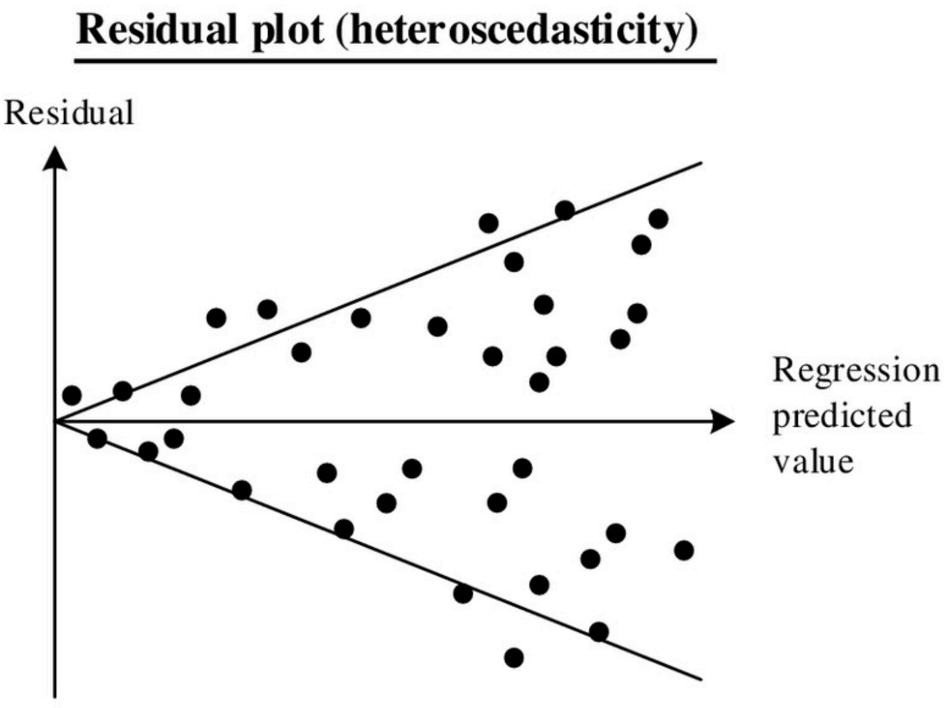


3.8 Residual Plots

(1) Plot of residuals versus predicted values.
(2) Plot of residuals versus explanatory value



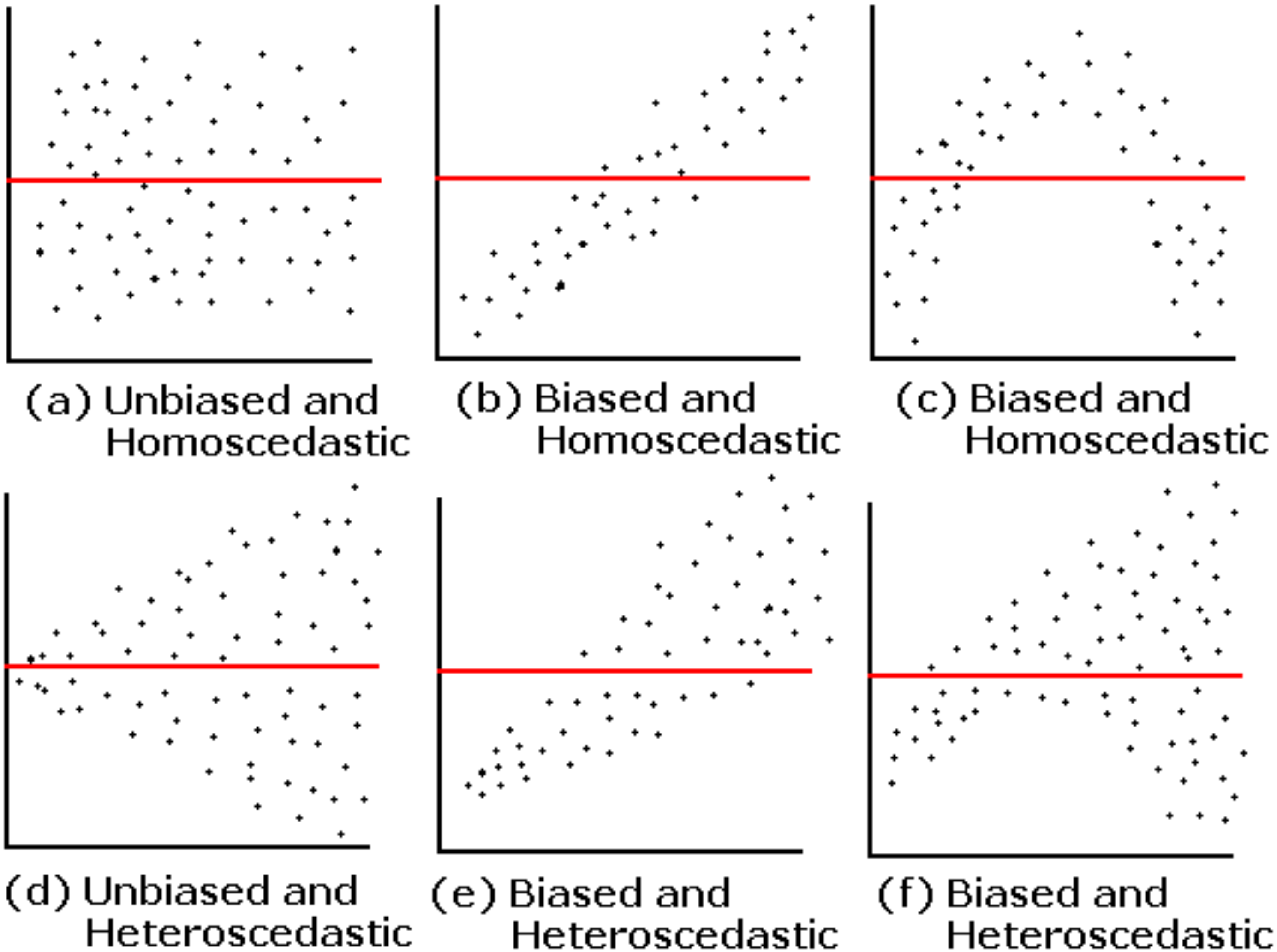
(a)



(b)

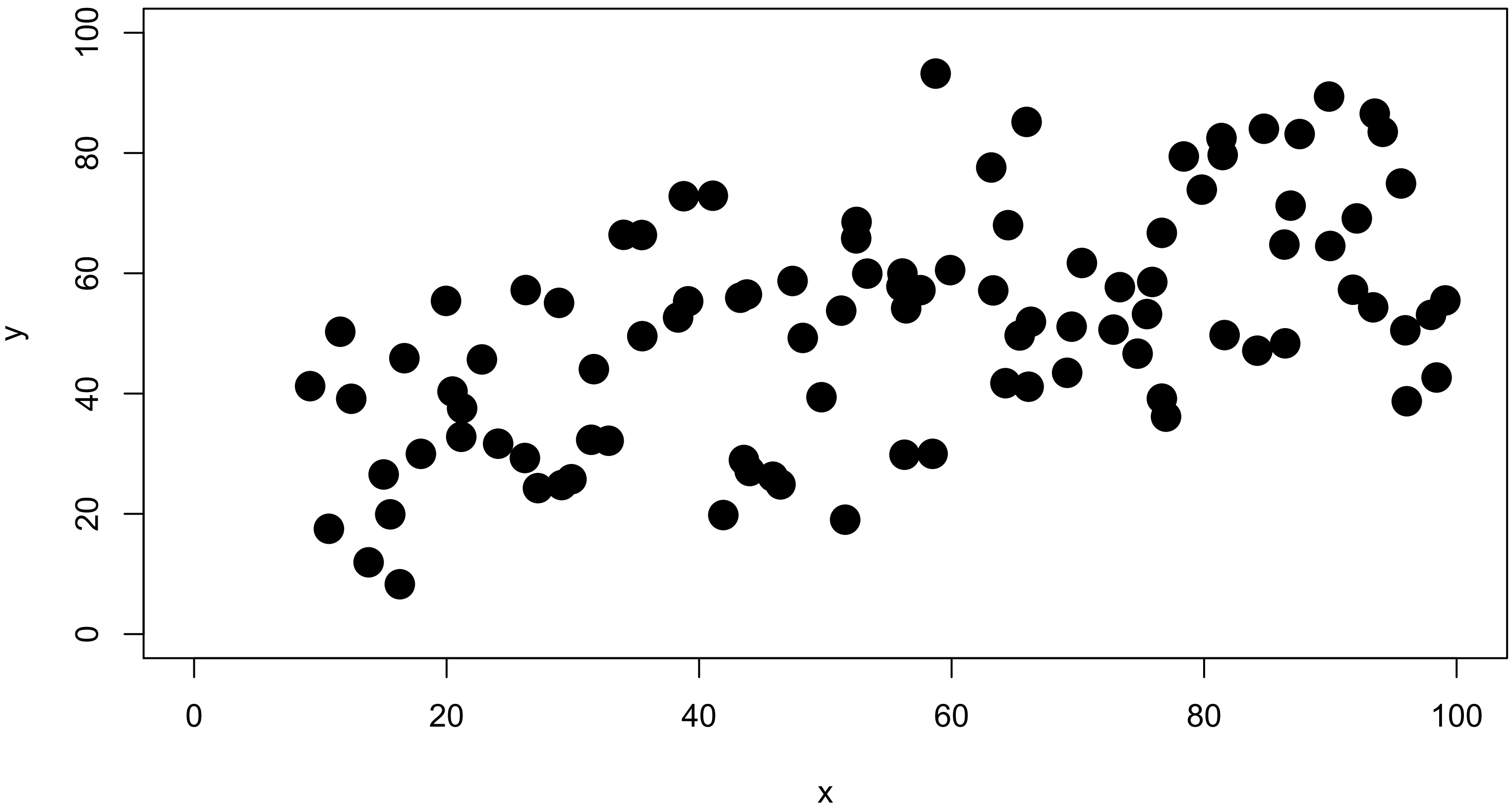
3.8 Residual Plots

(1) Plot of residuals versus predicted values.
(2) Plot of residuals versus explanatory value



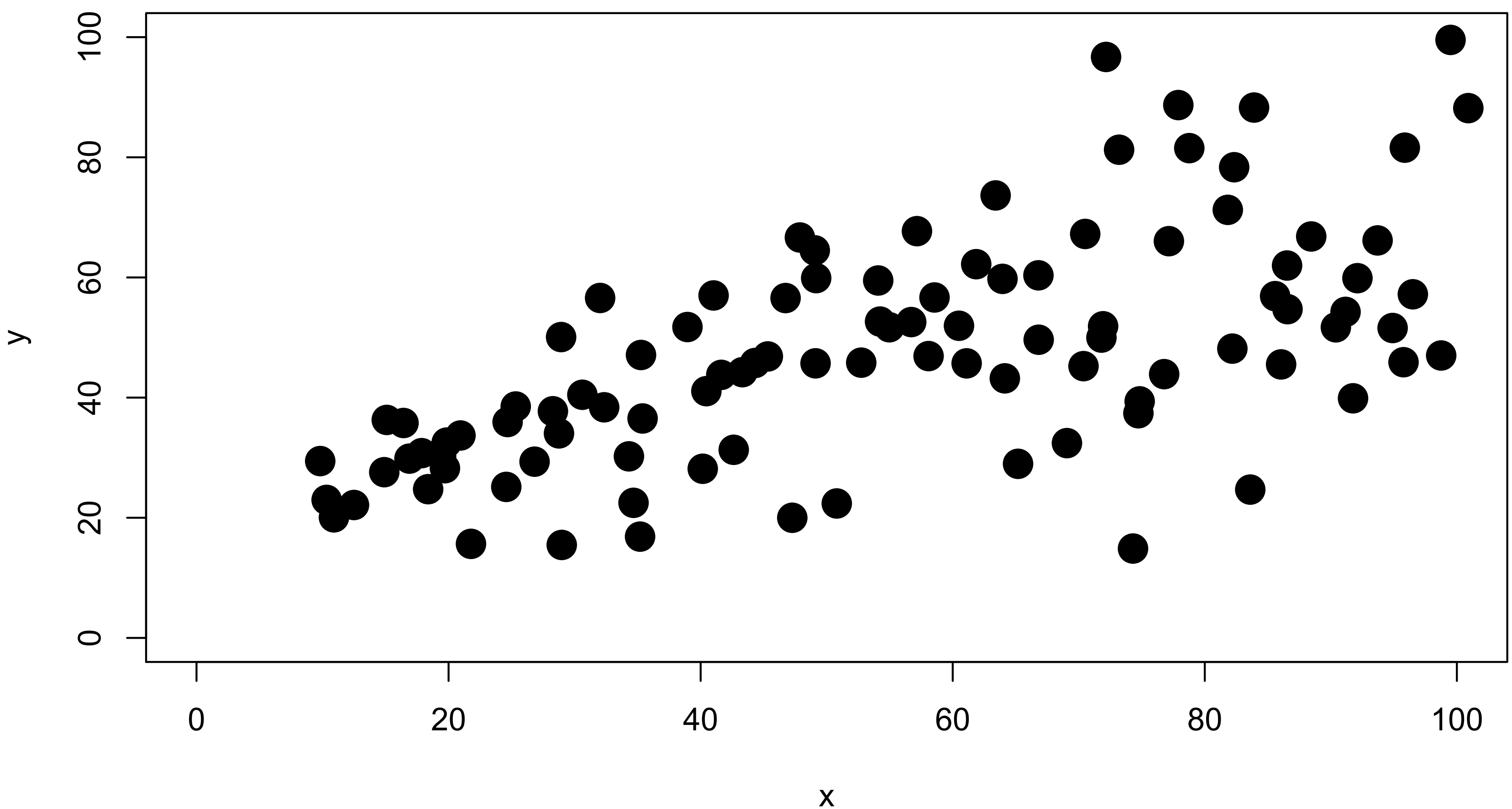
3.8 Residual Plots

Homoscedasticity of residuals or "equal variance"



3.8 Residual Plots

Heteroscedasticity of residuals or “not equal variance”



3.8 Residual Plots

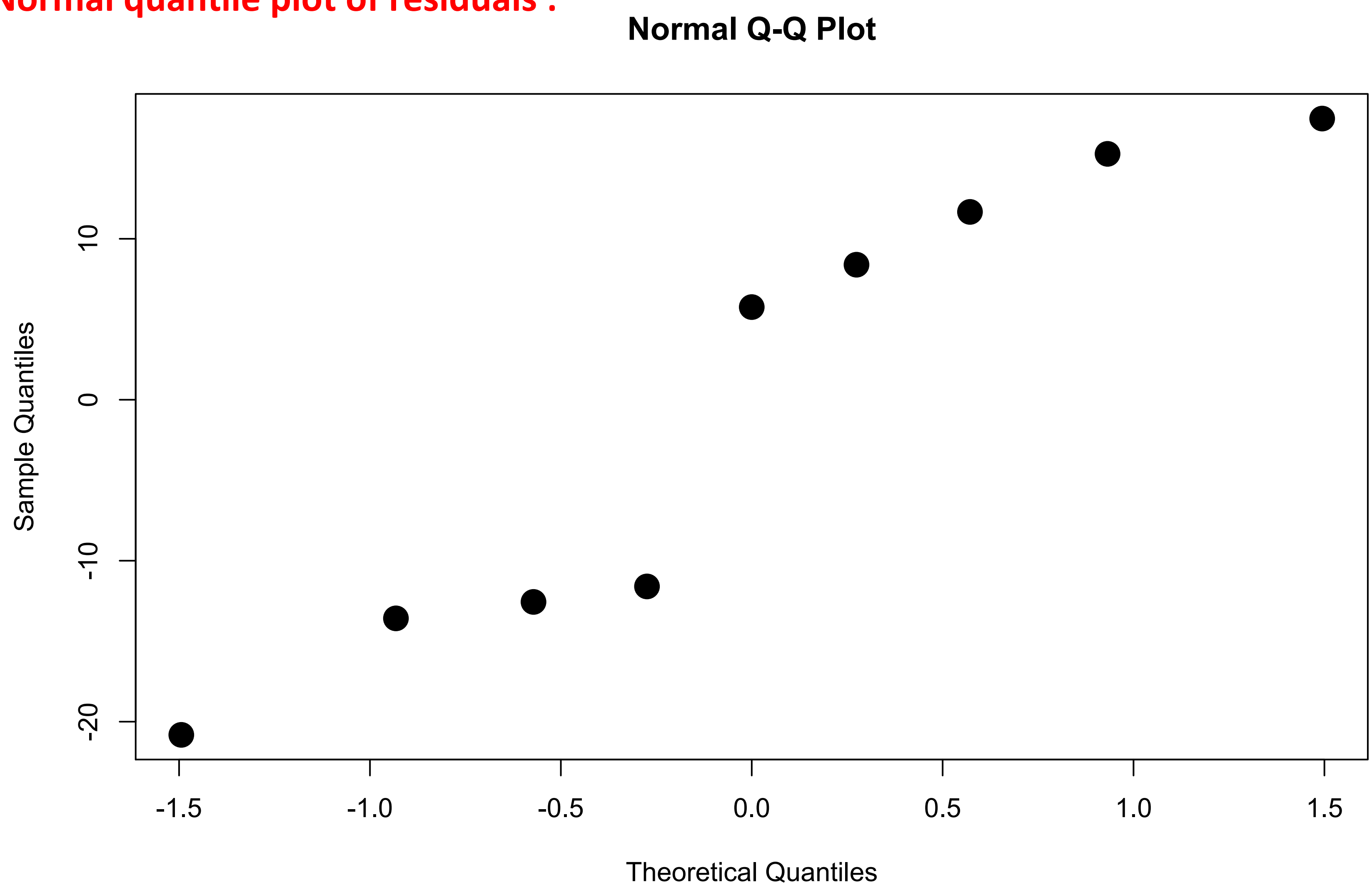
The response variable
is normally distributed.

(3) Normal quantile plot of residuals.

3.8 Residual Plots

The response variable is normally distributed.

(3) Normal quantile plot of residuals :



3.8 Residual Plots

(i) Check for homoscedasticity versus heteroscedasticity and possible structural deviations from model (plot of residuals versus predicted values, plots of residuals versus each explanatory variable).

(1) (2)

(ii) Check for normality (normal quantile plot of residuals) if the plots from (i) look OK.

(3)

now back to Chapter 2....

2.1.3 … Residuals

Because the best fitting line goes through the middle of the scatter of points, some e_i are ≥ 0 and others are ≤ 0. It turns out there is some balance and

$$\sum_{i=1}^{n} e_i = 0, \tag{2.29}$$

$$\bar{e} = n^{-1} \sum_{i=1}^{n} e_i = 0. \tag{2.30}$$

```
> sum(residuals)
[1] 1.332268e-14
> (1/n)*sum(residuals)
[1] 1.480297e-15
```

Chapter 2

- **Section 2.1** has the mathematics leading to the least squares line.

- **Section 2.2** introduces the simple linear regression model (prediction with one explanatory variable) that is formulated for a predictive equation. This is needed to quantify the variability of the coefficients of the best-fitting line, when different samples are taken from the population.

- **Section 2.5** has intervals for simple linear regression: the confidence interval for the slope of the least square line, confidence intervals for subpopulation means, and prediction intervals for a future or out-of-sample Y given x*.

- **Section 2.6** has an explanation of Student t quantiles used in the interval estimates.

Section 2.2 - Statistical linear regression model

1. For $i = 1, \ldots, n$, (x_i, y_i) is a realization of (x_i, Y_i), where Y_i is a random variable and x_i is non-random.
2. The stochastic relationship is:

   $$Y_i = \beta_0 + \beta_1 x_i + \epsilon_i, \quad i = 1, \ldots, n, \tag{2.32}$$

   where the ϵ_i's are independent normal random variables with mean 0 and variance σ^2. Think of ϵ as the sum of unmeasured effects.
3. From properties of normal random variables, this implies that

   $$Y_i \sim N(\beta_0 + \beta_1 x_i, \sigma^2). \tag{2.33}$$

Section 2.2 - Statistical linear regression model

1. For $i = 1, \ldots, n$, (x_i, y_i) is a realization of (x_i, Y_i), where Y_i is a random variable and x_i is non-random.
2. The stochastic relationship is:

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

   where the ϵ_i's are independent normal random variables with mean 0 and variance σ^2. Think of ϵ as the sum of unmeasured effects.
3. From properties of normal random variables, this implies that

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

Section 2.2 - Statistical linear regression model

1. For $i = 1, \ldots, n$, (x_i, y_i) is a realization of (x_i, Y_i), where Y_i is a random variable and x_i is non-random.
2. The stochastic relationship is:

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

   where the ϵ_i's are independent normal random variables with mean 0 and variance σ^2. Think of ϵ as the sum of unmeasured effects.
3. From properties of normal random variables, this implies that

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

What is a random variable?

Section 2.2 - Statistical linear regression model

1. For $i = 1, \ldots, n$, (x_i, y_i) is a realization of (x_i, Y_i), where Y_i is a random variable and x_i is non-random.
2. The stochastic relationship is:

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

   where the ϵ_i's are independent normal random variables with mean 0 and variance σ^2. Think of ϵ as the sum of unmeasured effects.
3. From properties of normal random variables, this implies that

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

What is a random variable?

A quantity for which it is impossible to know with 100% certainty it's value.

Section 2.2 - Statistical linear regression model

1. For $i = 1, \ldots, n$, (x_i, y_i) is a realization of (x_i, Y_i), where Y_i is a random variable and x_i is non-random.
2. The stochastic relationship is:

$$Y_i = \beta_0 + \beta_1 x_i + \epsilon_i, \quad i = 1, \ldots, n, \tag{2.32}$$

   where the ϵ_i's are independent normal random variables with mean 0 and variance σ^2. Think of ϵ as the sum of unmeasured effects.
3. From properties of normal random variables, this implies that

$$Y_i \sim N(\beta_0 + \beta_1 x_i, \sigma^2). \tag{2.33}$$

What is a random variable?

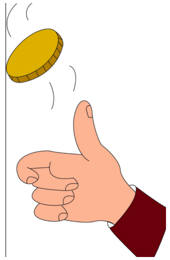
"A random variable, Y, is a variable whose possible values are numerical outcomes of a random phenomenon."

Section 2.2 - Statistical linear regression model

Three Examples of a Random Variable

"A random variable, Y, is a variable whose possible values are numerical outcomes of a random phenomenon."

Example 1: Y is the unknown result of a rolling a die (1, 2, 3, 4, 5 or 6)



Example 2: Y is the unknown result of a coin flip ("heads" or "tails")

Example 3: Y is the unknown amount of money that a random person walking on the street has in their bank account.

Section 2.2 - Statistical linear regression model

"A random variable, Y, is a variable whose possible values are numerical outcomes of a random phenomenon."

Example 1: Y is the **unknown** result of rolling a die (1, 2, 3, 4, 5 or 6)

Y is a random variable. All we can know about Y is that:

$Pr(Y=1) = 1/6$
$Pr(Y=2) = 1/6$
$Pr(Y=3) = 1/6$
$Pr(Y=4) = 1/6$
$Pr(Y=5) = 1/6$
$Pr(Y=6) = 1/6$

Section 2.2 - Statistical linear regression model

"A random variable, Y, is a variable whose possible values are numerical outcomes of a random phenomenon."

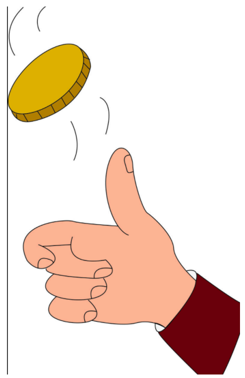
Example 2: Y is the **unknown** result of a coin flip ("heads" or "tails")

Once we observe the result of the coin flip we have "y".
"y" is not random variable it is a "realization of a random variable" also known as "data".

Y is a random variable that follows a Bernoulli(θ) distribution:

$$Y \sim \text{Bern}(\theta)$$

where θ is a population parameter.
For a "fair coin", θ = 0.5



Section 2.2 - Statistical linear regression model

"A random variable, Y, is a variable whose possible values are numerical outcomes of a random phenomenon."

Example 3: Y is the **unknown** amount of money that a random person has in their bank account.

Y is a random variable that may depend on the age (X) of the random person.

Let us assume that this dependence is linear such that:

$$Y = \beta_0 + \beta_1 X + \varepsilon \qquad \text{and} \qquad \varepsilon \sim \text{Normal}\,(0, \sigma^2)$$

where β_0 , β_1, and σ^2 are population parameters.

Section 2.2 - Statistical linear regression model

Y is a random variable that may depend on X.

We assume that this dependence is linear such that:

$$Y = \beta_0 + \beta_1 X + \varepsilon \qquad \text{and} \qquad \varepsilon \sim \text{Normal}\ (0, \sigma^2)$$

where β_0 , β_1, and σ^2 are population parameters.

1. For $i = 1, \ldots, n$, (x_i, y_i) is a realization of (x_i, Y_i), where Y_i is a random variable and x_i is non-random.
2. The stochastic relationship is:

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

   where the ϵ_i's are independent normal random variables with mean 0 and variance σ^2. Think of ϵ as the sum of unmeasured effects.
3. From properties of normal random variables, this implies that

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

Section 2.2 - Statistical linear regression model

Y is a random variable that may depend on X.

We assume that this dependence is linear such that:

$$Y = \beta_0 + \beta_1 X + \varepsilon \qquad \text{and} \qquad \varepsilon \sim \text{Normal}(0, \sigma^2)$$

where β_0, β_1, and σ^2 are population parameters.

ε is not a parameter. ε is a "random variable".

We have n random variables. For *i* = 1, …, n :

$$Y_i \sim \text{Normal}(\beta_0 + \beta_1 X_i, \sigma^2)$$

Section 2.2 - Statistical linear regression model

3. From properties of normal random variables, this implies that

$$(2.33) \qquad Y_i \sim N(\beta_0 + \beta_1 x_i,\, \sigma^2).$$

Conventional notation for probability/statistics

- Y (upper case letter) is a random variable;
- y (lower case letter) is a realization of a random variable;
- boldfaced $\mathbf{Y} = \begin{pmatrix} Y_1 \\ Y_2 \\ \vdots \\ Y_n \end{pmatrix}$ is a random vector and $\mathbf{y} = \begin{pmatrix} y_1 \\ y_2 \\ \vdots \\ y_n \end{pmatrix}$ is a vector of realized values;

Section 2.2 - Statistical linear regression model

3. From properties of normal random variables, this implies that

$$Y_i \sim N(\beta_0 + \beta_1 x_i, \sigma^2). \tag{2.33}$$

Conventional notation for probability/statistics

- Y (upper case letter) is a random variable;
- y (lower case letter) is a realization of a random variable;
- boldfaced $\mathbf{Y} = \begin{pmatrix} Y_1 \\ Y_2 \\ \vdots \\ Y_n \end{pmatrix}$ is a random vector and $\mathbf{y} = \begin{pmatrix} y_1 \\ y_2 \\ \vdots \\ y_n \end{pmatrix}$ is a vector of realized values;
- inside $\mathrm{E}(\cdot)$ and $\mathrm{Var}(\cdot)$, the arguments are random variables and they should be shown in upper case;

Section 2.2 - Statistical linear regression model

3. From properties of normal random variables, this implies that

$$Y_i \sim N(\beta_0 + \beta_1 x_i,\, \sigma^2). \tag{2.33}$$

Conventional notation for probability/statistics

- Y (upper case letter) is a random variable;
- y (lower case letter) is a realization of a random variable;
- boldfaced $\mathbf{Y} = \begin{pmatrix} Y_1 \\ Y_2 \\ \vdots \\ Y_n \end{pmatrix}$ is a random vector and $\mathbf{y} = \begin{pmatrix} y_1 \\ y_2 \\ \vdots \\ y_n \end{pmatrix}$ is a vector of realized values;
- inside $\mathrm{E}(\cdot)$ and $\mathrm{Var}(\cdot)$, the arguments are random variables and they should be shown in upper case;
- Greek letters are used for parameters such as β_1, σ, μ (an exception is the use of ϵ for the random deviation from the line or curve);
- caret or hat on a symbol is used for estimators, for example $\hat{\mu}, \hat{\sigma}$ (read as mu hat or sigma hat).

Section 2.2 - Statistical linear regression model

What is a random variable?

"A random variable, Y, is a variable whose possible values are numerical outcomes of a random phenomenon."

For a Random Variable, Y, we typically want to talk about the Expectation and Variance:

Example 1: $E[Y] = 3.5$ $\quad$ $Var(Y) = 2.92$

Example 2: $E[Y] = 0.5$ $\quad$ $Var(Y) = 0.25$

Example 3: $E[Y] = \beta_0 + \beta_1 X$ $\quad$ $Var(Y) = \sigma^2$

Questions?

Section 2.2 - Statistical linear regression model

Example 1: All we can know about Y is that:

$Pr(Y=1) = 1/6$ $Pr(Y=2) = 1/6$ $Pr(Y=3) = 1/6$

$Pr(Y=4) = 1/6$ $Pr(Y=5) = 1/6$ $Pr(Y=6) = 1/6$

```
> Y<-sample(c(1,2,3,4,5,6), size=1,prob=c(1/6,1/6,1/6,1/6,1/6,1/6))
```

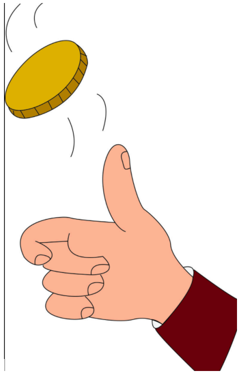
Using the definitions of Expectation and Variance we can calculate:

$E[Y] = 3.5$ $Var(Y) = 2.92$

or:

```
> mean(sample(c(1,2,3,4,5,6), size=10000000,prob=c(1/6,1/6,1/6,1/6,1/6,1/6), replace=TRUE)
[1] 3.500276
> var(sample(c(1,2,3,4,5,6), size=10000000,prob=c(1/6,1/6,1/6,1/6,1/6,1/6), replace=TRUE))
[1] 2.918512
>
```

Section 2.2 - Statistical linear regression model



Example 2: All we can know about Y is that:

$Pr(Y=0) = 1-\theta$

$Pr(Y=1) = \theta$

```
> theta<-0.5
> Y<-sample(c(0,1), 1, prob=c(1-theta, theta))
```

Using the definitions of Expectation and Variance we can calculate:

$E[Y] = \theta$ $\quad\quad$ $Var[Y] = \theta\,(1-\theta)$

or:

```
> mean(sample(c(0,1), 100000, prob=c(1-theta,
theta),replace=TRUE))
[1] 0.50282
```

Section 2.2 - Statistical linear regression model

Example 3 for a fixed value of X:

Let X=20.

$Y = \beta_0 + \beta_1 20 + \varepsilon$ and $\varepsilon \sim \text{Normal}(0, \sigma^2)$

Therefore: $Y \sim \text{Normal}(\beta_0 + \beta_1 20 , \sigma^2)$

Using properties of the Normal distribution:

$E[Y] = \beta_0 + \beta_1 20$ $\quad$ $Var(Y) = \sigma^2$

```
> beta0<-20
> beta1<-0.5
> X<-20
> sigma2<-100
> Y<-rnorm(n=1, mean=beta0+beta1*X, sd=sqrt(sigma2))
```

Section 2.2 - Statistical linear regression model

Example 3 for a fixed value of X:

Y ~ Normal ($\beta_0 + \beta_1 20$, σ^2)

```
> beta0<-20
> beta1<-0.5
> X<-20
> sigma2<-100
> Y<-rnorm(n=1, mean=beta0+beta1*X, sd=sqrt(sigma2))
```

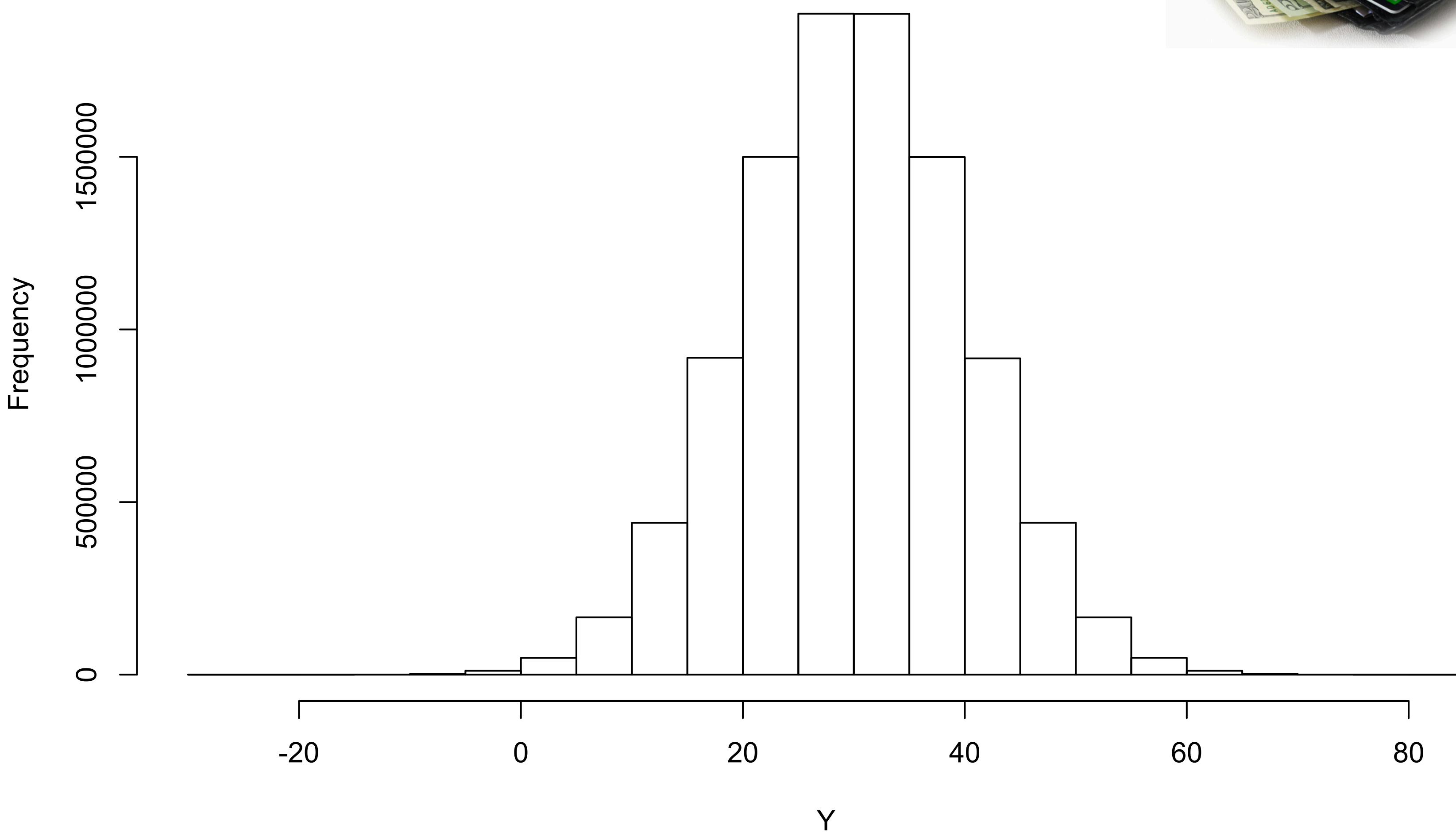
Let's take a large sample of Y and look at the distribution with a histrogram:

```
Y<-rnorm(n=10000000, mean=beta0+beta1*X, sd=sqrt(sigma2))
hist(Y)
```

Section 2.2 - Statistical linear regression model

Example 3 for a fixed value of X=20:

Y ~ Normal ($\beta_0 + \beta_1 20$, σ^2)

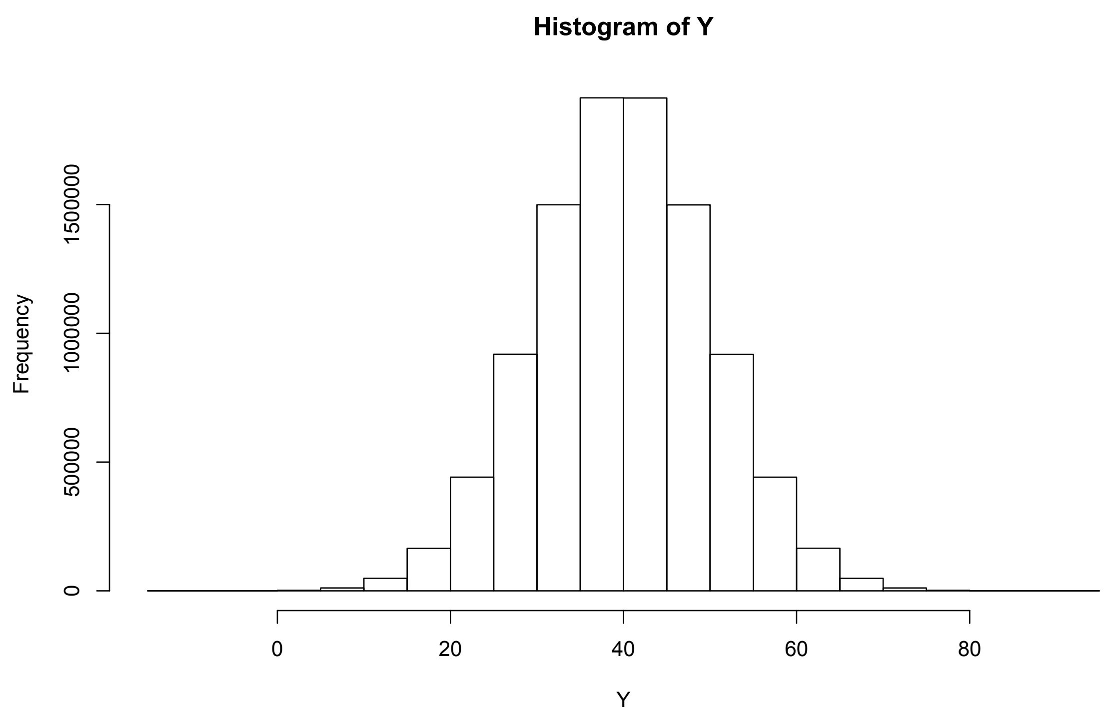


Section 2.2 - Statistical linear regression model

Example 3 for a fixed value of X=40:

$X = 40$, $Y \sim \text{Normal}(\beta_0 + \beta_1 40 , \sigma^2)$

```
> X<-40
> Y<-rnorm(n=10000000, mean=beta0+beta1*X, sd=sqrt(sigma2))
> hist(Y)
```

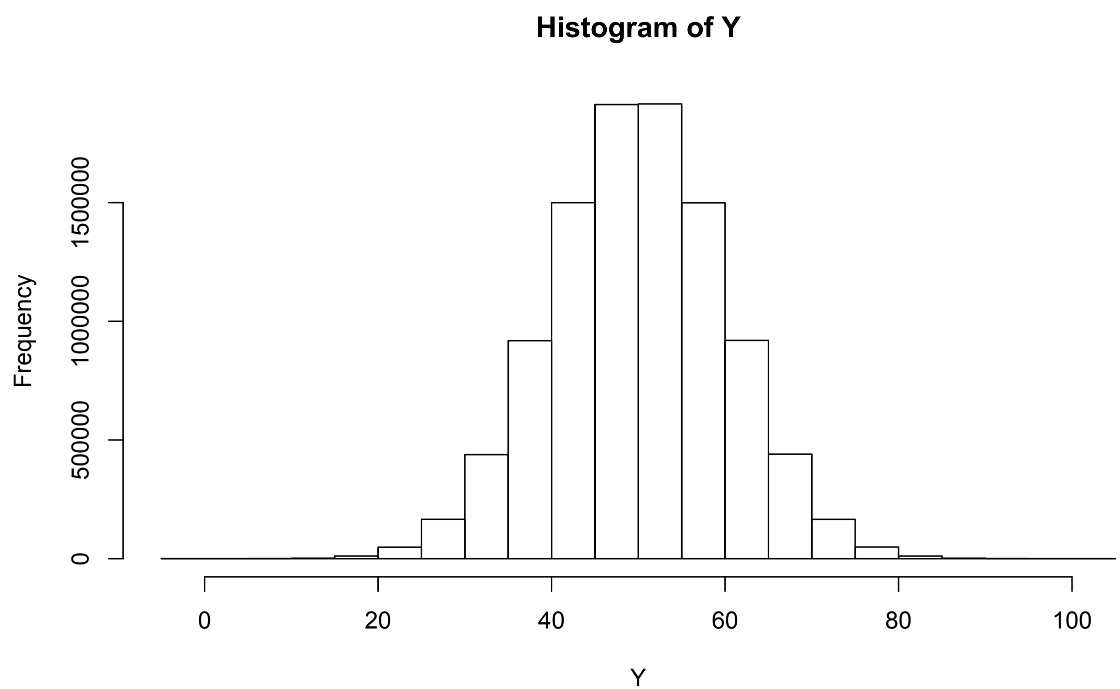


Section 2.2 - Statistical linear regression model

Example 3 for a fixed value of X=60:

$X = 60$, $Y \sim \text{Normal}(\beta_0 + \beta_1 60 , \sigma^2)$

```
> X<-60
> Y<-rnorm(n=10000000, mean=beta0+beta1*X, sd=sqrt(sigma2))
> hist(Y)
```

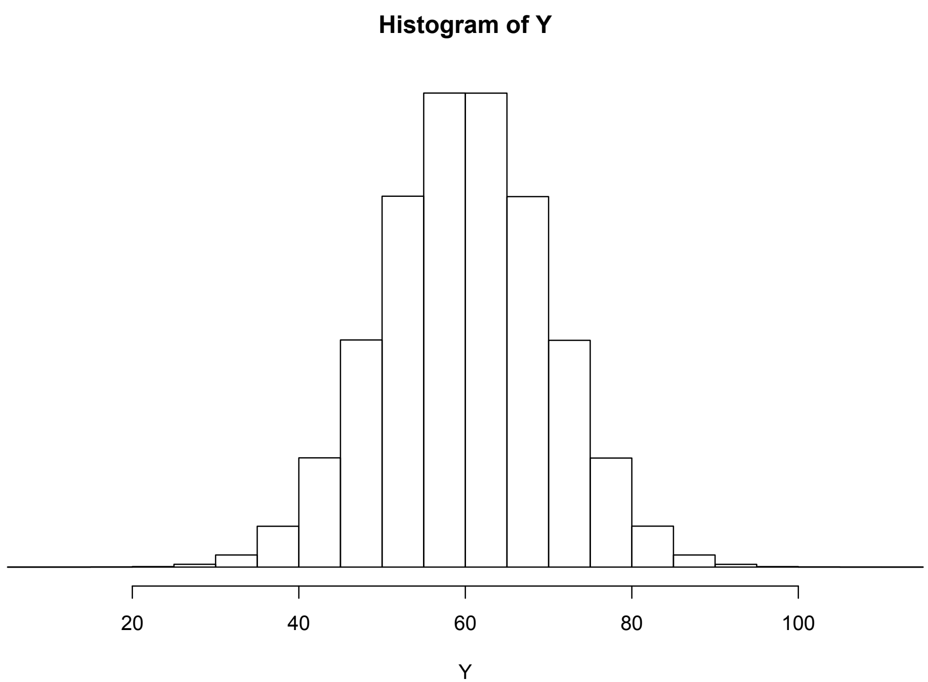


Section 2.2 - Statistical linear regression model

Example 3 for a fixed value of X=80:

$X = 80$, $Y \sim \text{Normal}(\beta_0 + \beta_1 80 , \sigma^2)$

```
> X<-80
> Y<-rnorm(n=10000000, mean=beta0+beta1*X, sd=sqrt(sigma2))
> hist(Y)
>
```



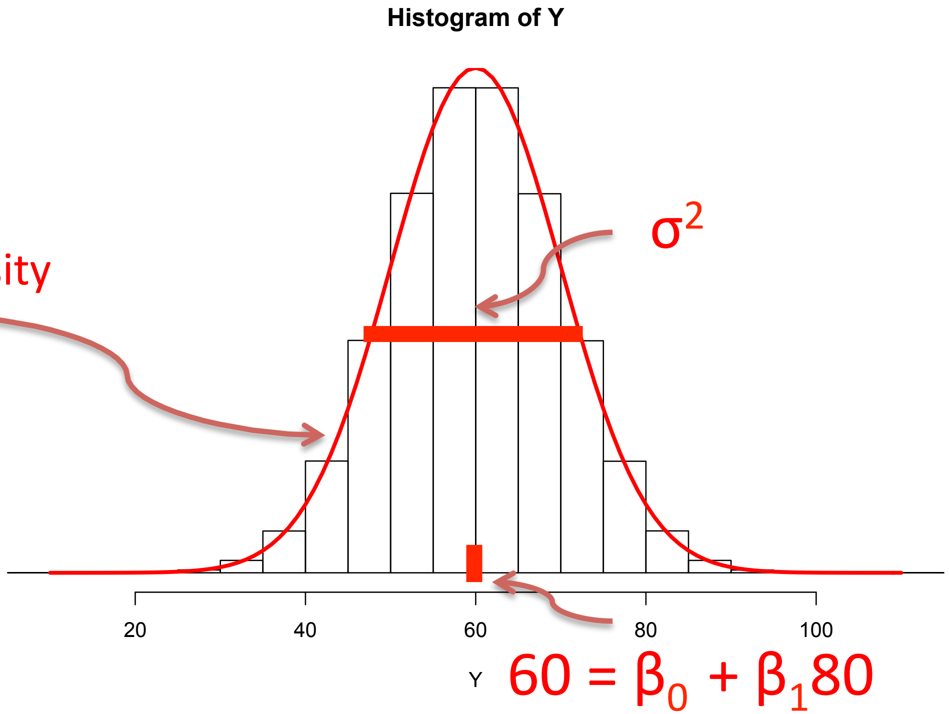
Section 2.2 - Statistical linear regression model

Example 3 for a fixed value of X=80:

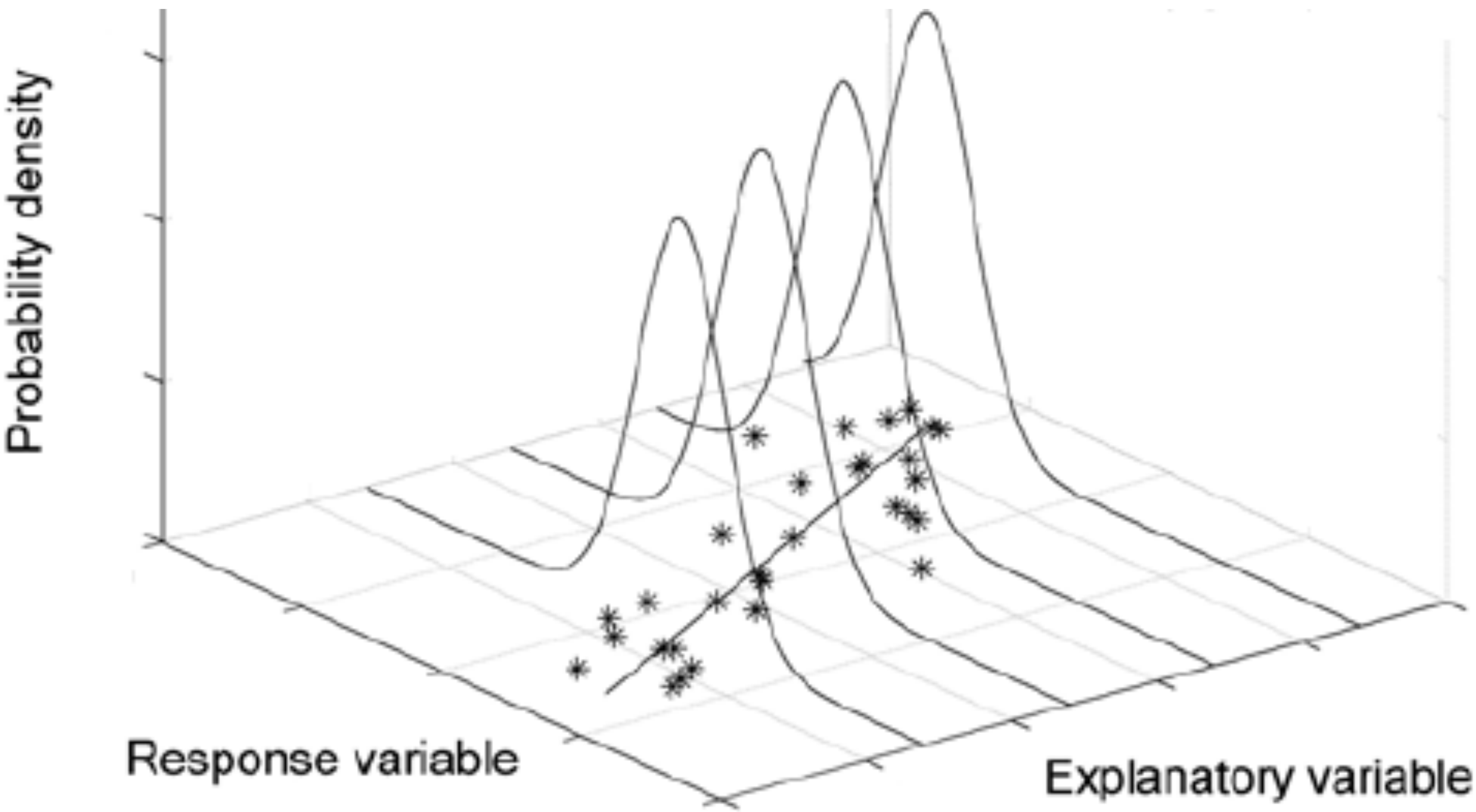
$X = 80$, $Y \sim \text{Normal}(\beta_0 + \beta_1 80 , \sigma^2)$

```
> X<-80
> Y<-rnorm(n=10000000, mean=beta0+beta1*X, sd=sqrt(sigma2))
> hist(Y)
>
```

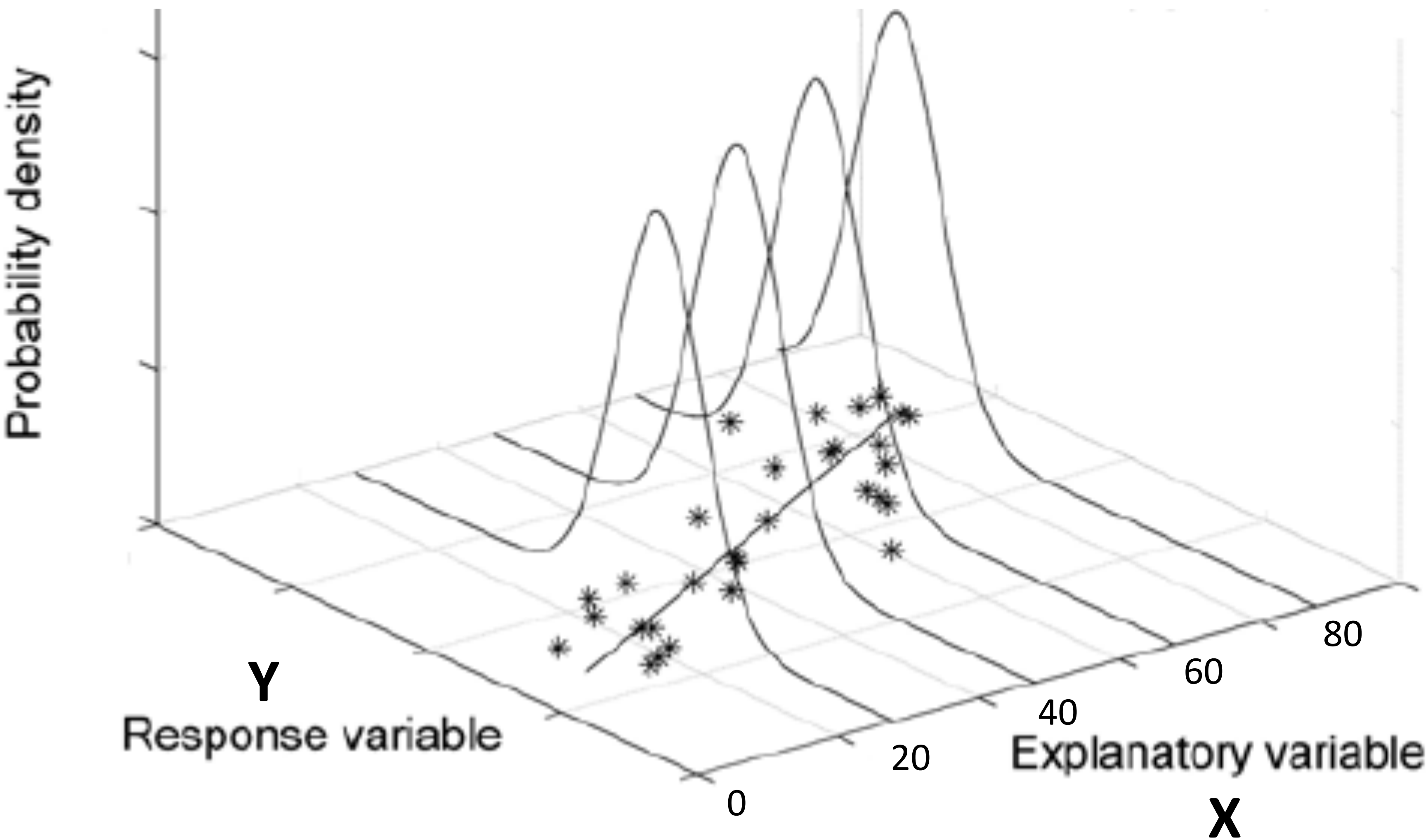
probability density



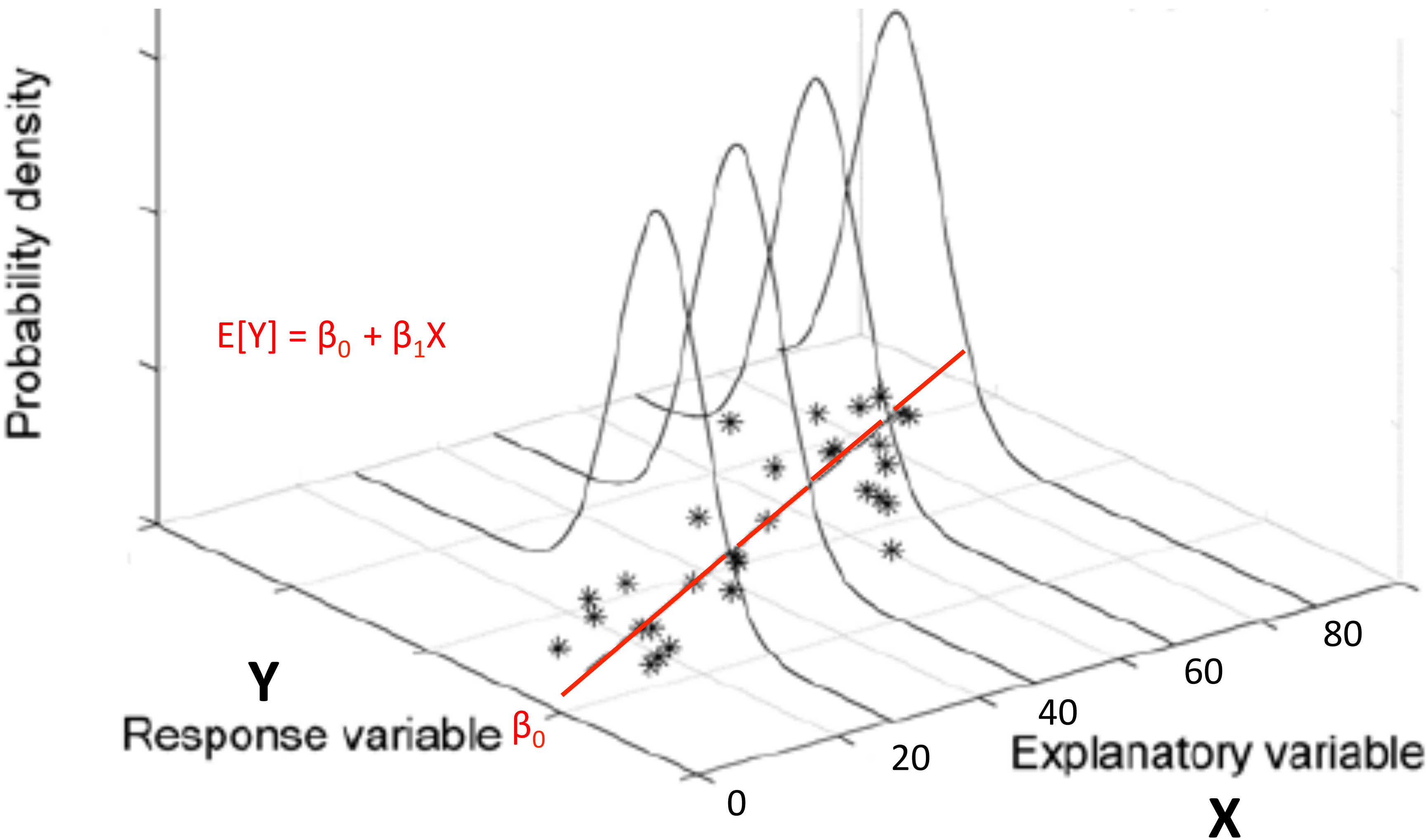
Section 2.2 - Statistical linear regression model



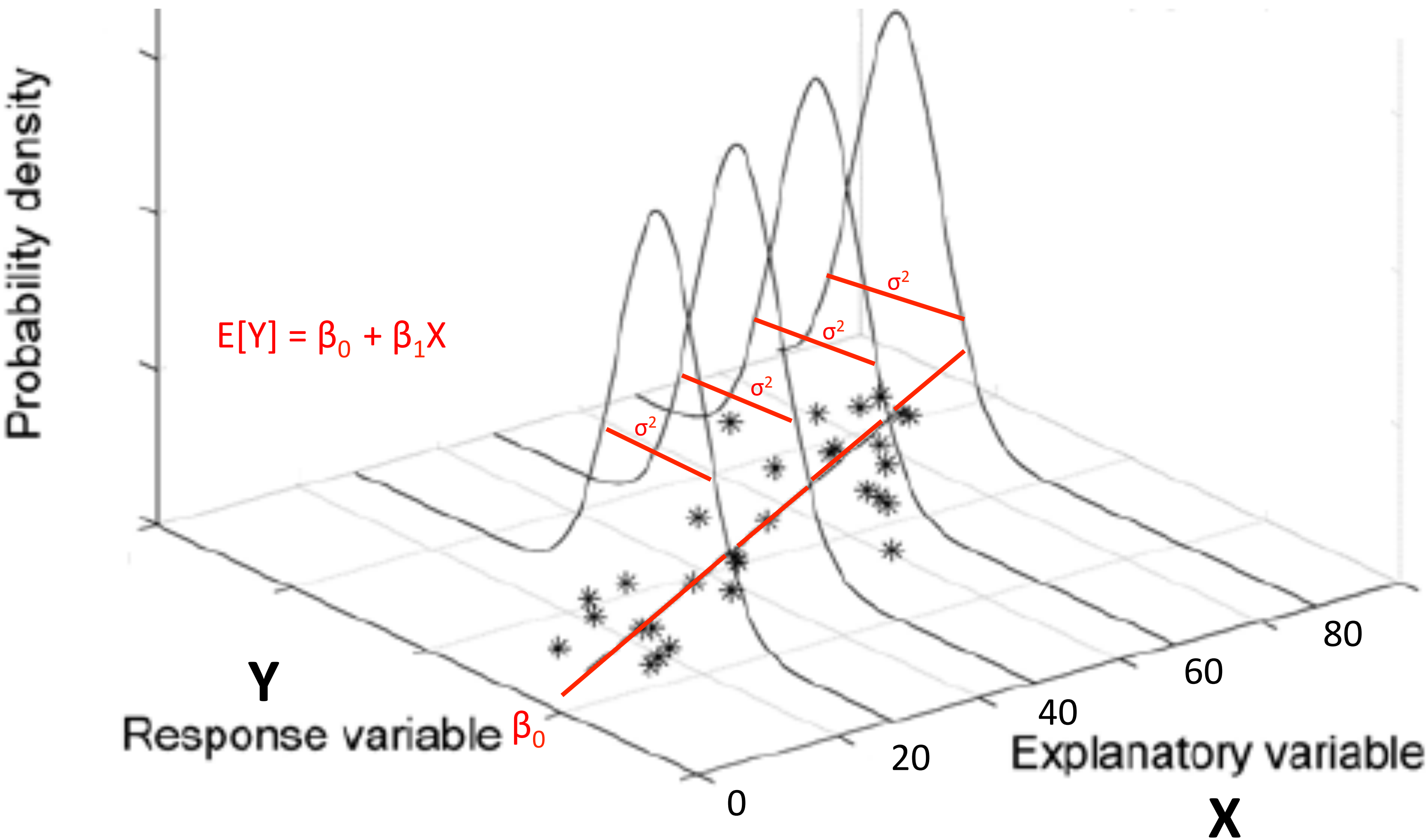
Section 2.2 - Statistical linear regression model



Section 2.2 - Statistical linear regression model

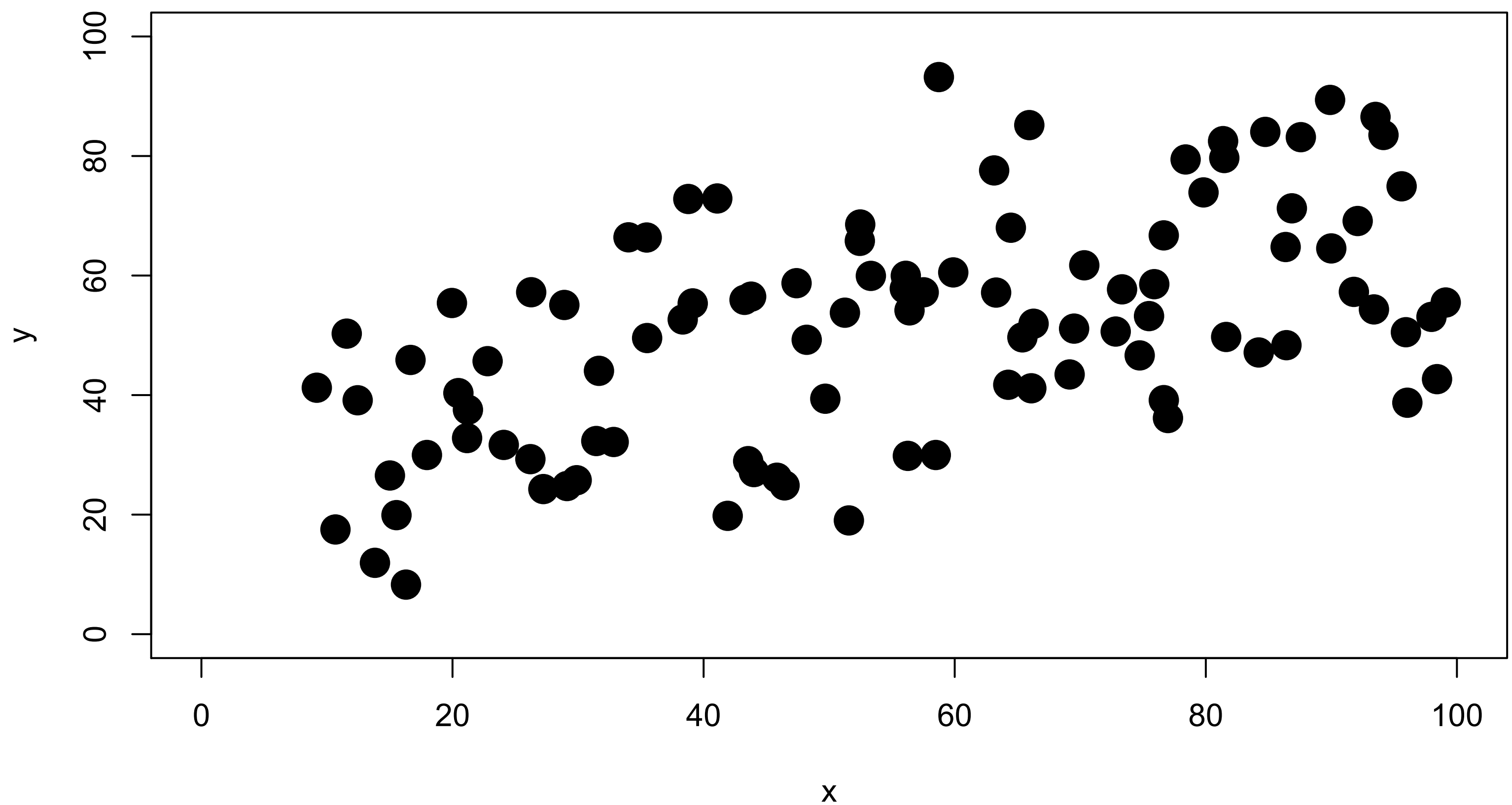


Section 2.2 - Statistical linear regression model



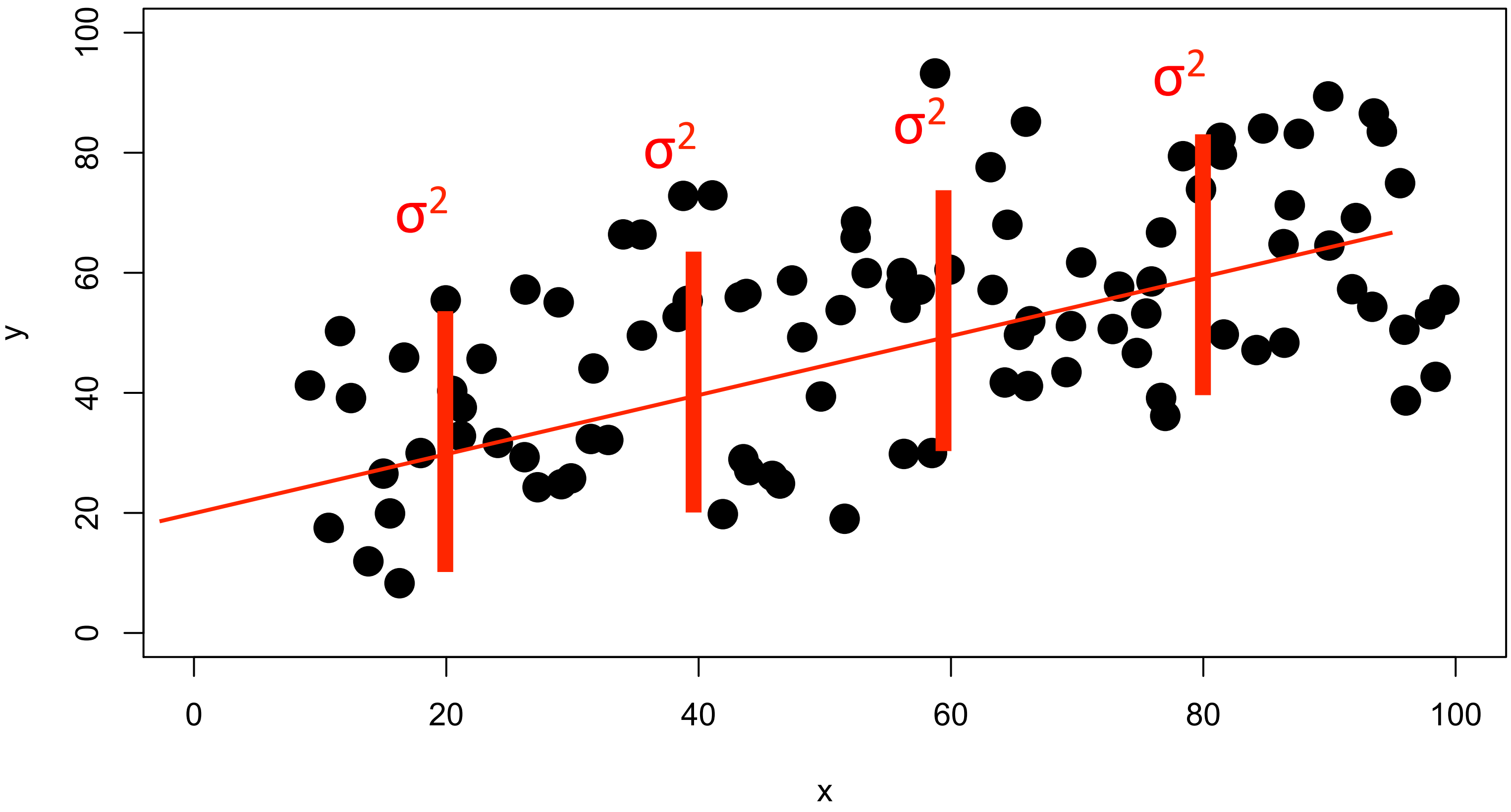
3.8 Residual Plots

Homoscedasticity of residuals or “equal variance”



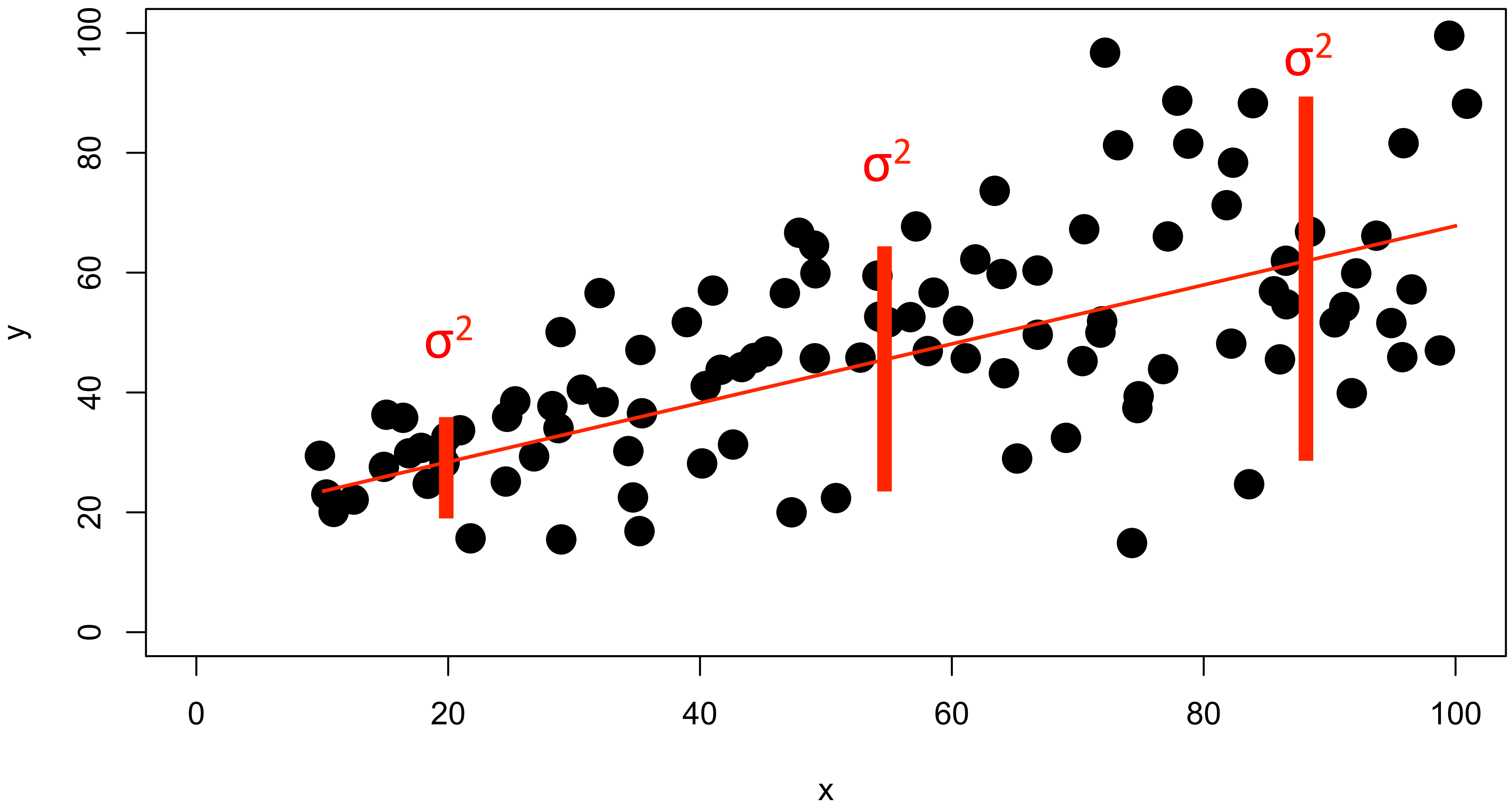
3.8 Residual Plots

Homoscedasticity of residuals or "equal variance"



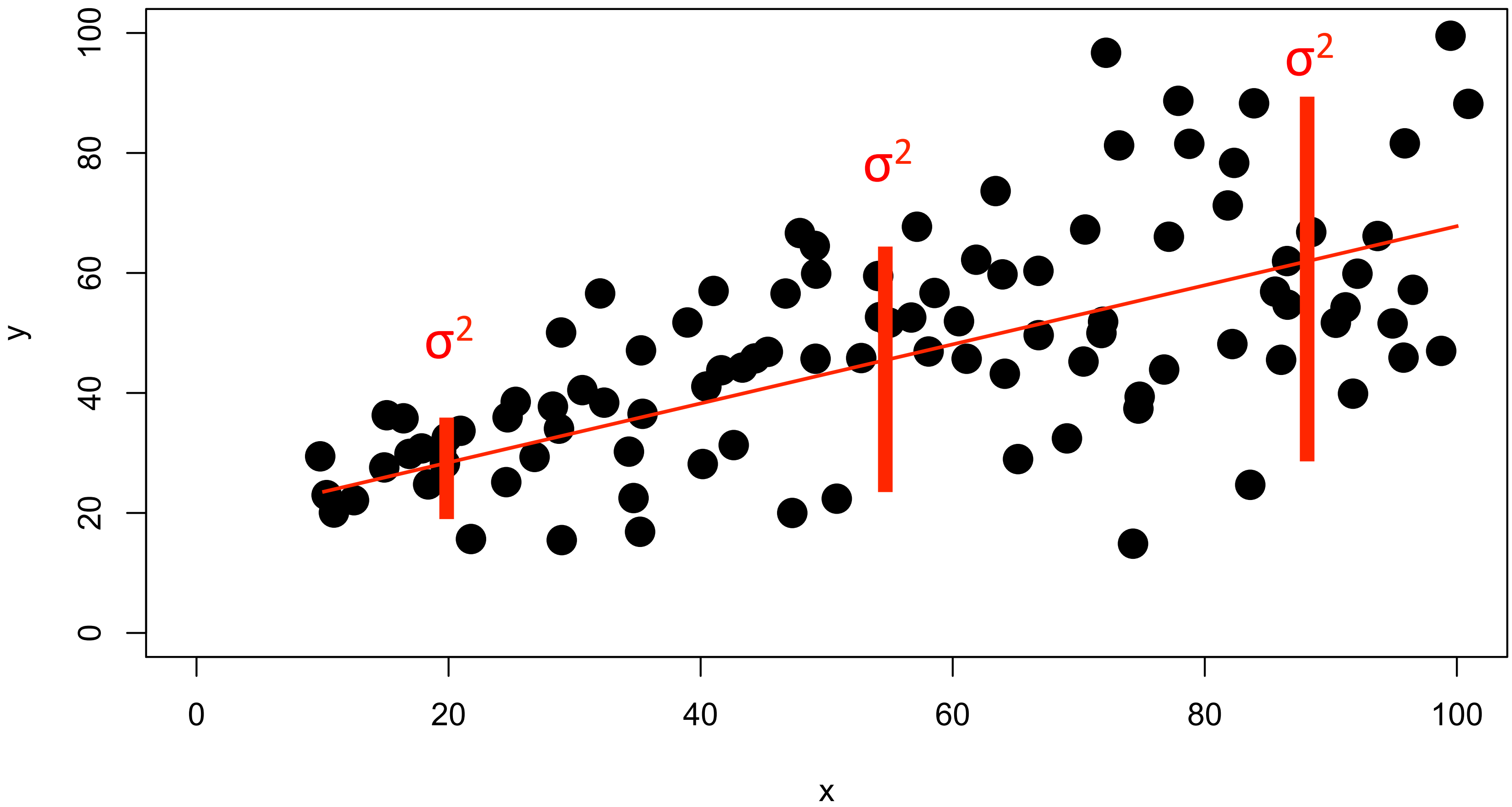
3.8 Residual Plots

Heteroscedasticity of residuals or “not equal variance”



3.8 Residual Plots

Heteroscedasticity of residuals or “not equal variance”



Questions?