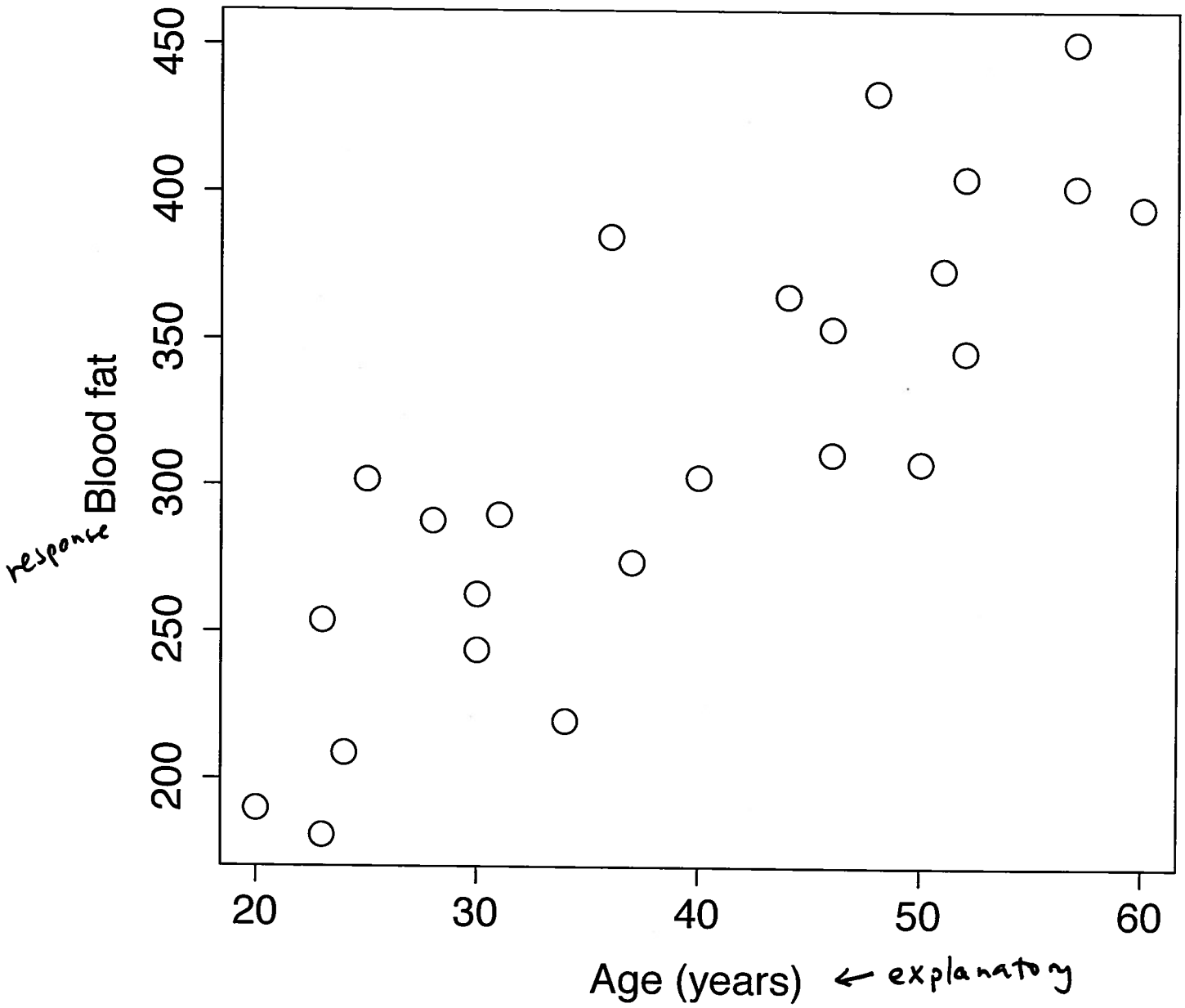
# Chapter 11 Simple Linear Regression

Objective: to describe a linear relationship between two quantitative variables using a model. The model fits a straight line to the data and can be used to make predictions on the **response variable** ($y$) given the **explanatory variable** ($x$).

For example, the intensity ($y$) of strong motion earthquakes is dependent on the distance from the epicenter ($x$). We may be interested in estimating the relationship between $y$ and the explanatory variable based on some sample.

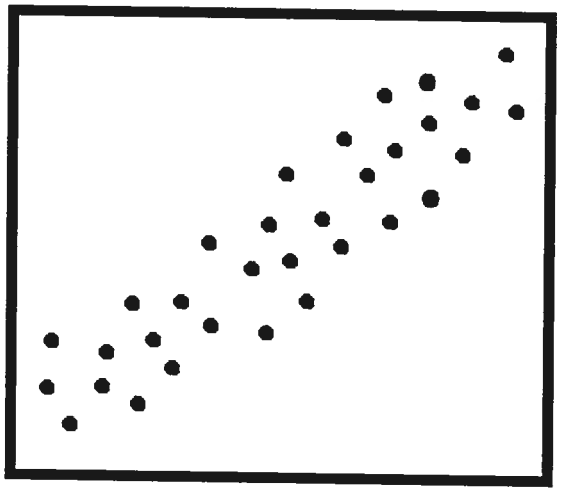
# Motivating Example

Let's examine blood fat data for 25 individuals aged between 20 to 60 years (D.G. Kleinbaum and L.L. Kupper, *Applied Regression Analysis and Other Multivariable Methods*). Blood fat can be used as a measurement for assessing cardiovascular health and high levels of blood fat in the body can lead to heart issues. Simple blood tests can be used to determine the blood fat level in a patient.

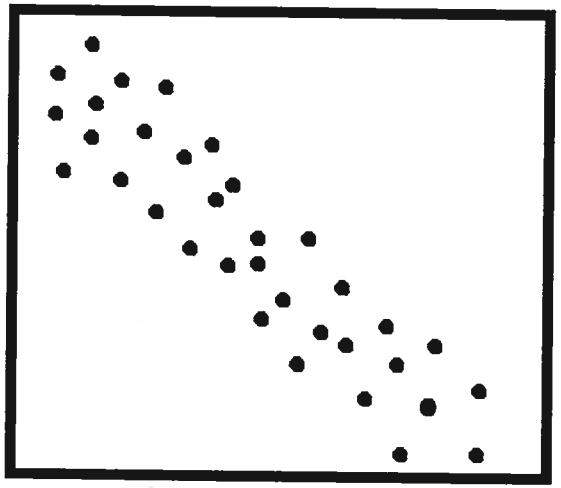

Blood fat
response
200
250
300
350
400
450
20
30
40
50
60
Age (years)
explanatory

When we look at scatterplots we will look for:

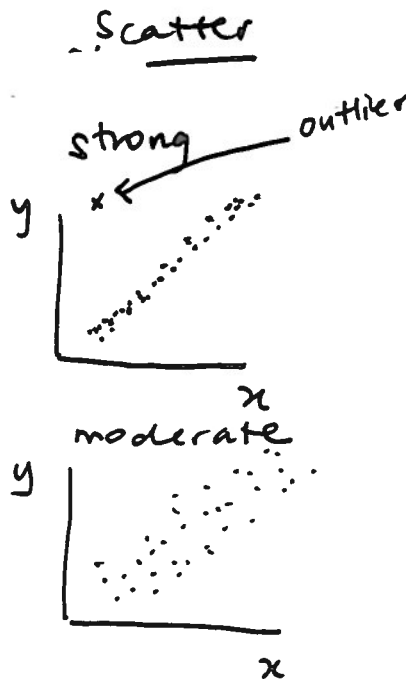
1. Direction (positive, negative, none)
   - 1.1 If large values of $x$ are associated with large values of $y$, or as $x$ increases the corresponding value of $y$ tends to increase, then there is a **positive relationship** between $x$ and $y$.
   - 1.2 If small values of $x$ are associated with large values of $y$, or as $x$ increases the corresponding value of $y$ tends to decrease, then there is a **negative relationship** between $x$ and $y$.
2. Form (linear, curved, no clear form)
3. Scatter (strong relationship, weak or no relationship)
4. Any outliers?

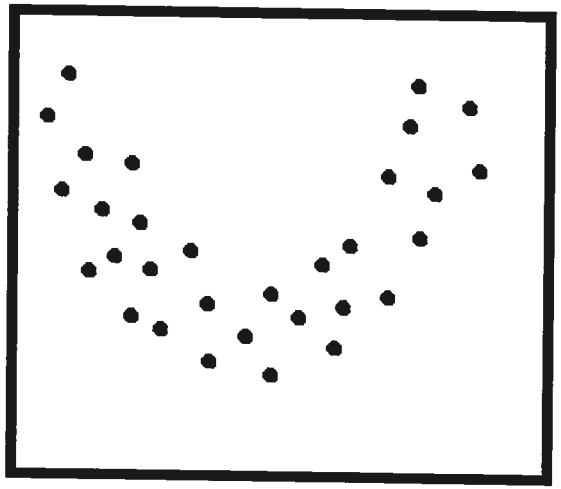


positive linear association

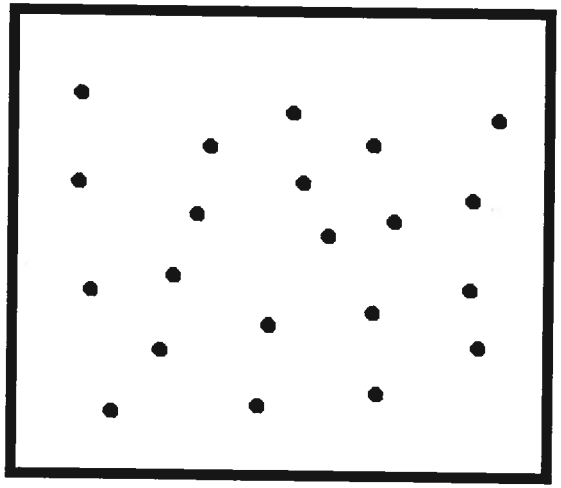


negative linear association



nonlinear association



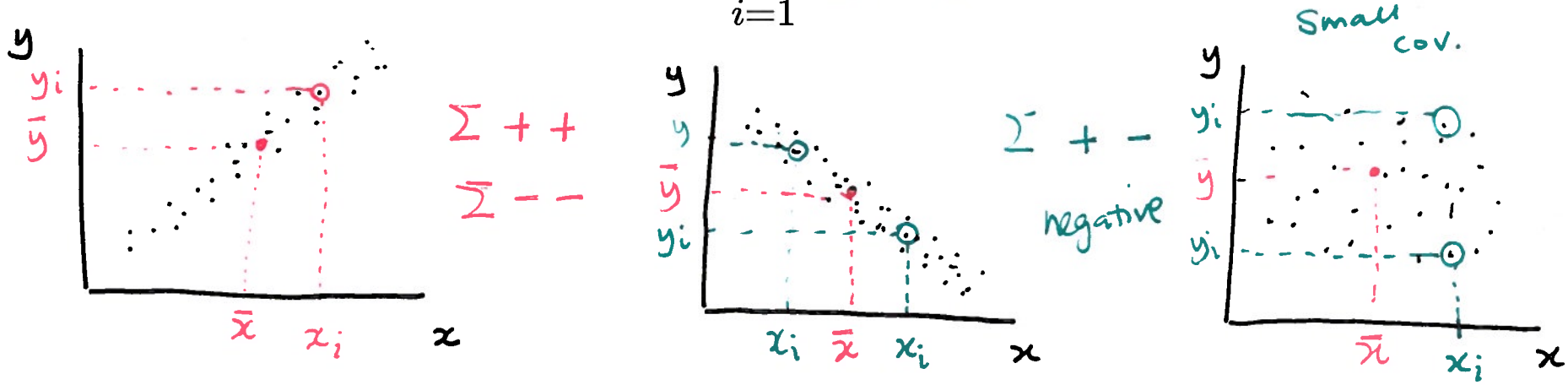
no association

# Covariance and Correlation Coefficient (ch. 2)

We can quantify the degree of linear association between pairs of variables using the **covariance** and the **correlation coefficient** $(r)$.

The covariance is a measure of how much two random variables change together. The sample covariance is

$$Cov(x,y) = \frac{1}{n-1}\sum_{i=1}^{n}(x_i - \bar{x})(y_i - \bar{y})$$



# Covariance

- If $x$ and $y$ are positively associated then $Cov(x, y)$ will be large and positive
- If $x$ and $y$ are negatively associated then $Cov(x, y)$ will be large and negative
- If the variables are not positively nor negatively associated then $Cov(x, y)$ will be small

The **correlation coefficient** gives us a numerical measurement of the strength of the linear relationship between two quantitative variables.

$$r = \frac{Cov(x,y)}{s_x s_y} = \frac{1}{n-1}\sum_{i=1}^{n}\left(\frac{x_i - \bar{x}}{s_x}\right)\left(\frac{y_i - \bar{y}}{s_y}\right)$$

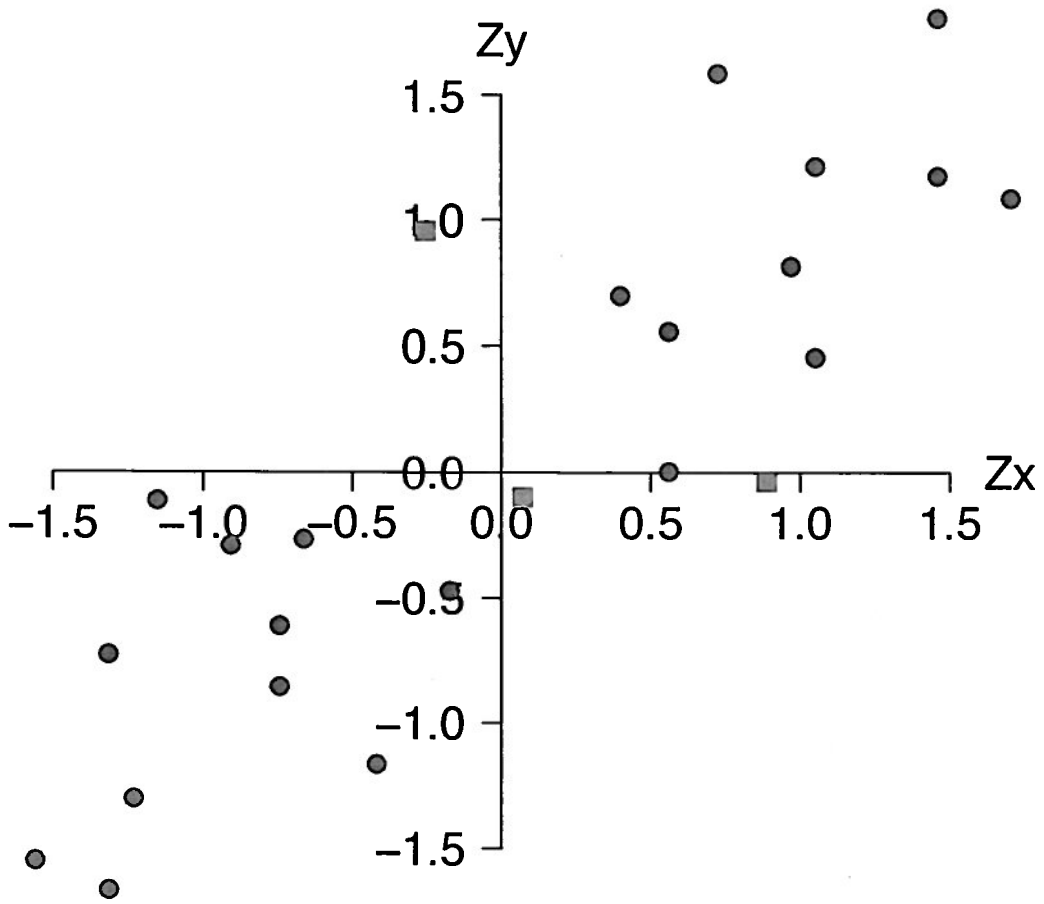
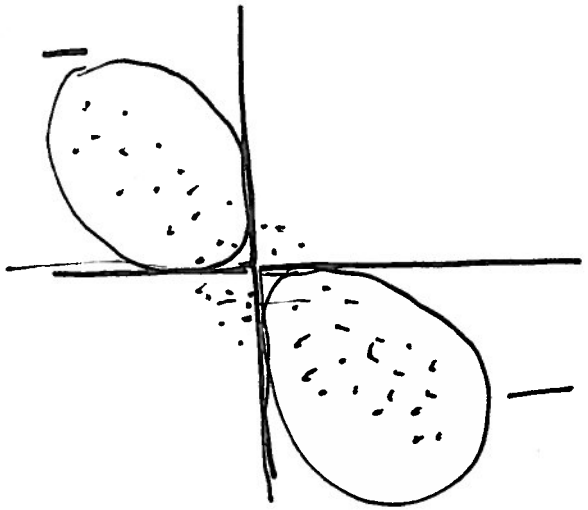
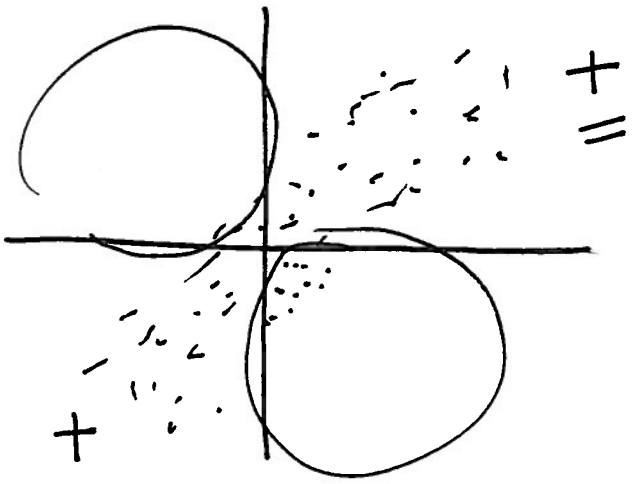
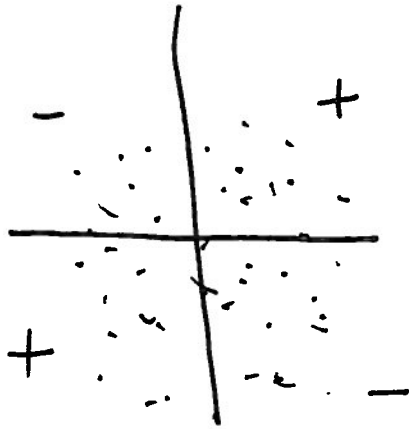


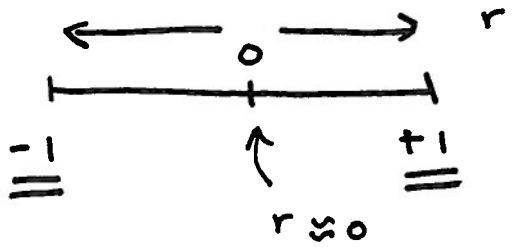

Figure 1: Scatterplot of the standardized blood fat and age. Both variables are standardized and the coordinates of a point are written as $(z_x, z_y)$. Some points (blue circles) strengthen the impression of a positive association between height and weight. Other points (red squares) tend to weaken the positive association.

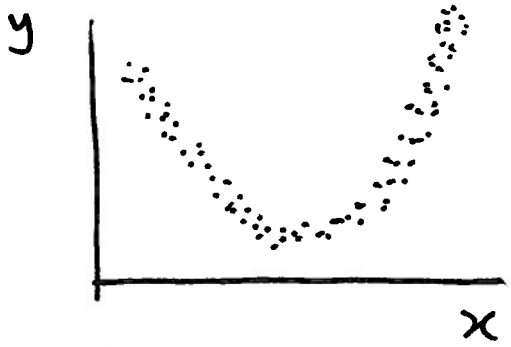



$-1 \leq r \leq +1$



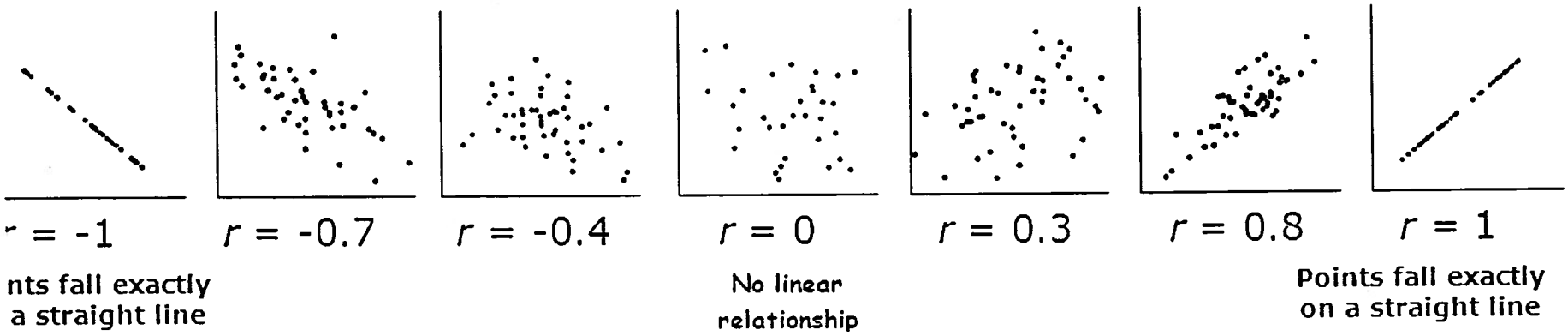




$r \approx 0$

## Properties of Correlation Coefficient

- $-1 \leq r \leq 1$ and has no units
- the sign of $r$ tells us the direction of the relationship
- degree of positive correlation increases: $r$ becomes closer to 1
  degree of negative correlation increases: $r$ becomes closer to -1
- $r$ close to 0 implies very weak or no linear relationship between the two variables (but this does not imply the two variables are not related in another way, e.g. non-linear relationship)

- $r$ has a value between -1 and +1:



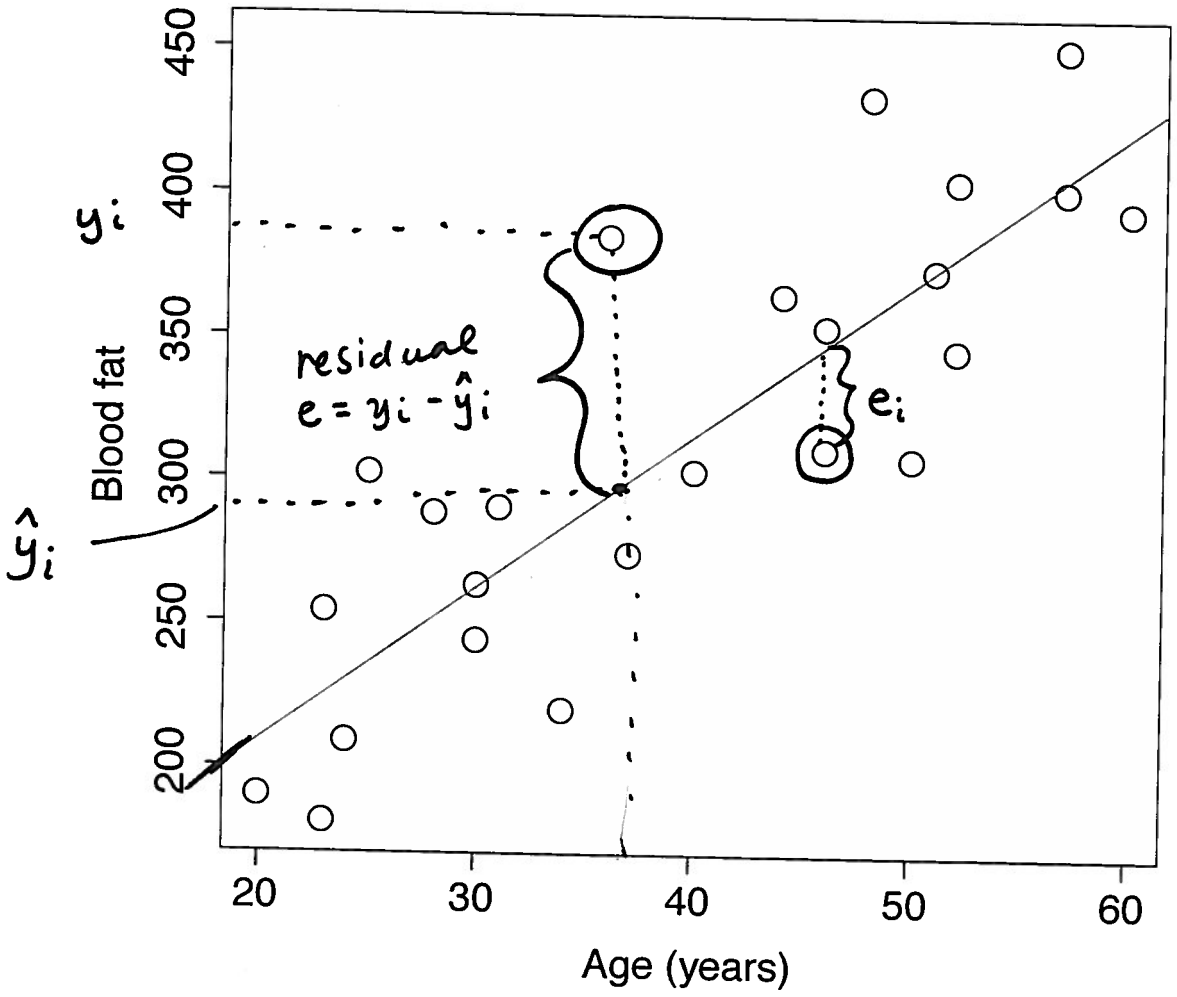
# Least Squares Method

The **regression line** is a line that best describes the relationship between $x$ and $y$. Linear regression consists of finding the best-fitting straight line (**line of best fit**) through the points. The line that best describes the relationship between $x$ and $y$ is:

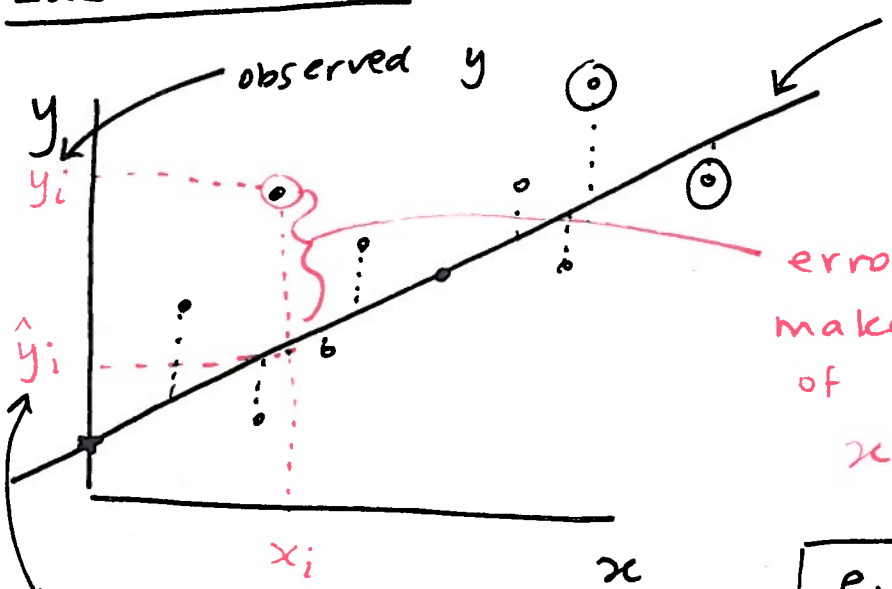
| Regression line ($y$ on $x$) |
| --- |
| $\hat{y} =$ intercept $+$ (slope) $\times$ $(x)$ |
| where $\hat{y}$ is the predicted value |

The value of $\hat{y}$ does not necessarily coincide with the value of $y$ for any given $x$. (The line does not necessarily pass through all the points).

For our data, we could fit a regression line through the points as such:



# Last Class:




model predicts certain y ($\hat{y}$)

line of best fit

$\hat{y}$ = intercept + slope x

what values give us best fitting line?

error our model would make in predicting y of that particular x.

$e_i = y_i - \hat{y}_i$ Residuals

prediction error.

Goal: minimize squared residual

$$\min \sum (y_i - \hat{y}_i)^2$$

$\widehat{\text{blood fat}} = 102.6 + 5.32\ \text{age}$

$b_0$ $b_1$

population parameters.

$b_0 \rightarrow \beta_0$

$b_1 \rightarrow \beta_1$

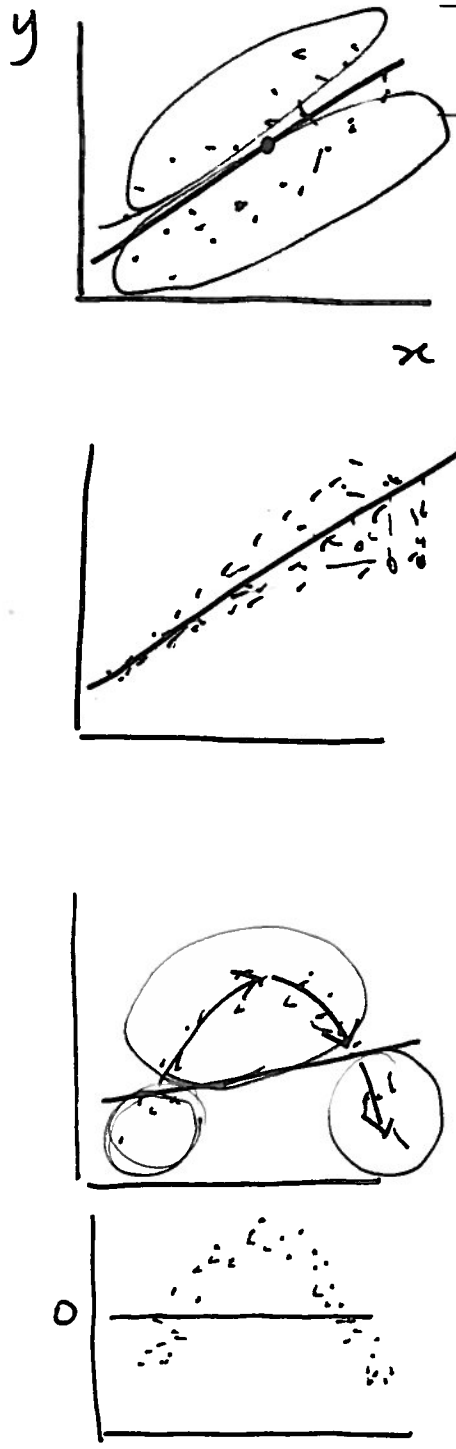
how useful is this model? $r \frac{S_y}{S_x} = 0.75$ response.

# Residuals

The residual ($e$) is the difference between the observed $y$ and the predicted value $\hat{y}$,

$$\boxed{e = y - \hat{y}}$$

- The linear model is obtained by minimizing the sum of the squared residuals (vertical distances from the observed points to the line). Thus we refer to the linear model as the **least squares regression line**.
- Aim: minimize $\sum_{i}^{n} e_i^2$ (see pg. 171 for proof)




| Subject | Age $x$ | Bloodfat $y$ | Predicted $\hat{y} = 102.6 + 5.32x$ | Residuals $e = y - \hat{y}$ |
|---|---|---|---|---|
| 1 | 46 | 354 | 347.33 | 6.67 |
| 2 | 20 | 190 | 208.99 | -18.99 |
| 3 | 52 | 405 | 379.25 | 25.75 |
| 4 | 30 | 263 | 262.20 | 0.80 |
| 5 | 57 | 451 | 405.85 | 45.15 |
| 6 | 25 | 302 | 235.59 | 66.41 |
| 7 | 28 | 288 | 251.55 | 36.45 |
| 8 | 36 | 385 | 294.12 | 90.88 |
| 9 | 57 | 402 | 405.85 | -3.85 |
| 10 | 44 | 365 | 336.68 | 28.32 |
| 11 | 24 | 209 | 230.27 | -21.27 |
| 12 | 31 | 290 | 267.52 | 22.48 |
| 13 | 52 | 346 | 379.25 | -33.25 |
| 14 | 23 | 254 | 224.95 | 29.05 |
| 15 | 60 | 395 | 421.82 | -26.82 |
| 16 | 48 | 434 | 357.97 | 76.03 |
| 17 | 34 | 220 | 283.48 | -63.48 |
| 18 | 51 | 374 | 373.93 | 0.07 |
| 19 | 50 | 308 | 368.61 | -60.61 |
| 20 | 34 | 220 | 283.48 | -63.48 |
| 21 | 46 | 311 | 347.33 | -36.33 |
| 22 | 23 | 181 | 224.95 | -43.95 |
| 23 | 37 | 274 | 299.44 | -25.44 |
| 24 | 40 | 303 | 315.40 | -12.40 |
| 25 | 30 | 244 | 262.20 | -18.20 |

354 - 347.33 = 6.67

residual plot

e

x

$b_1$

The least squares estimates of the regression line is

$$\hat{y} = b_0 + b_1 x$$

Slope:

$$b_1 = \frac{r s_y}{s_x} = \frac{n\sum x_i y_i - (\sum x_i)(\sum y_i)}{n\sum x_i^2 - (\sum x_i)^2}$$

where $s_x$ and $s_y$ are the standard deviations of $x$ and $y$ respectively and $r$ is correlation

Intercept:

$$b_0 = \bar{y} - b_1 \bar{x} = \frac{\sum y_i - b_1 \sum x_i}{n}$$

- The slope ("rise over run") tells us how much a change in $y$ to expect for a unit increase in $x$. If the slope is positive (negative), $y$ increases (decreases) with $x$.
- The intercept tells us the $y$ value when $x$ takes a value of $0$. It is the point that crosses the $y$-axis.
- The line passes through the mean-mean point $(\bar{x}, \bar{y})$

| Subject | Age | Fat |
|---|---|---|
| 1 | 46 | 354 |
| 2 | 20 | 190 |
| 3 | 52 | 405 |
| ⋮ | ⋮ | ⋮ |
| 25 | 30 | 244 |
| Mean | 39.12 | 310.7 |
| SD | 12.25 | 77.83 |
| r | 0.8373 | |

## Example 1

Using our blood fat example, what is $b_0$ and $b_1$?

Sol: $b_1 = \frac{rs_y}{s_x} = 0.8373 \times \frac{77.83}{12.25} = 5.32$

$b_0 = \bar{y} - b_1\bar{x} = 310.7 - 5.32 \times 39.12 = 102.6$

## Example 2

Predict the blood fat level for an 26 year old individual.

Sol: $\widehat{\text{blood fat}} = 102.6 + 5.32 \times \text{age} = 102.6 + 5.32 \times 26 = 240.92$

# Inference for Regression

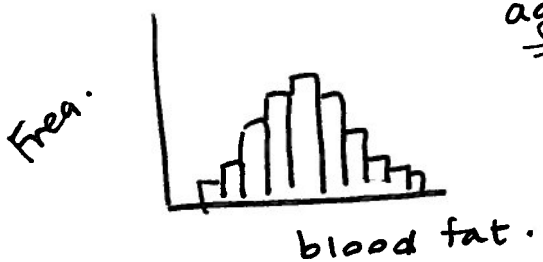
We can use inference to make conclusions about how the population of $y$ values relate to the population of $x$ values, based on the sample $x$ and sample $y$ values. The equation

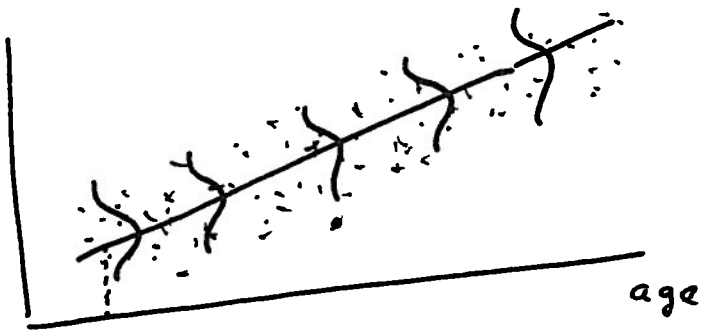
$$\mu_Y = \beta_0 + \beta_1 x$$

describes this population relationship. We use Greek letters to denote the **coefficients** (intercept and slope) since these are parameters.

→ CI and hypothesis test $\beta_1$

For each age




$$\mu_Y = \beta_0 + \beta_1 x$$

↑ parameter coefficients.

$$y = \beta_0 + \beta_1 x + \varepsilon$$

↑ random errors. positive, negative, zero.

$\hat{y} = b_0 + b_1 x$ estimate $\beta$'s

→ sampling distributions.

If we had all the values in the population, we could find the slope and intercept of this idealized regression line by using least squares. The line can't actually match all the values in the population. Some ys lie above or below the line so for each data point, the model makes an error. The errors are random and can be positive, negative or zero.

$$y = \beta_0 + \beta_1 x + \varepsilon$$

where $\varepsilon$ represents model errors

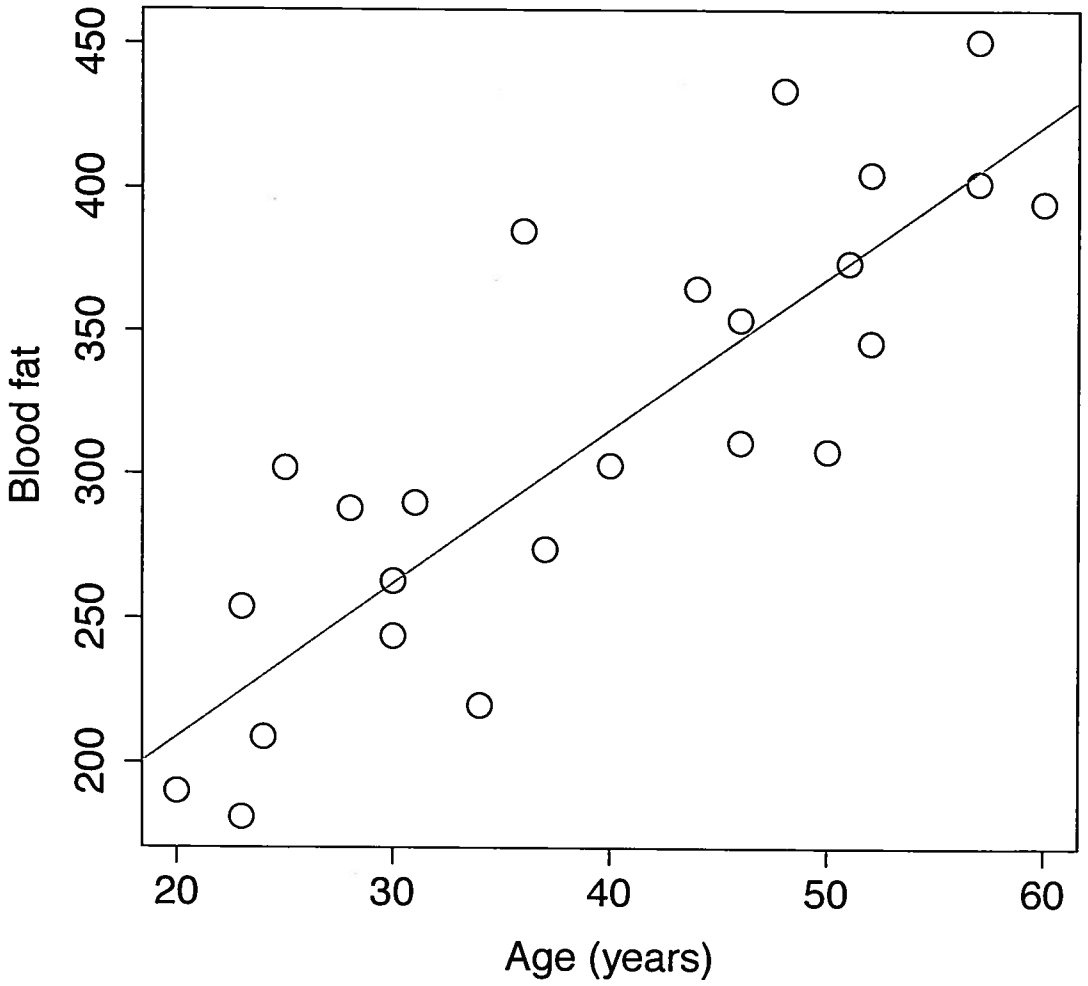
We estimate the true linear relationship with

$$\hat{y} = b_0 + b_1 x$$

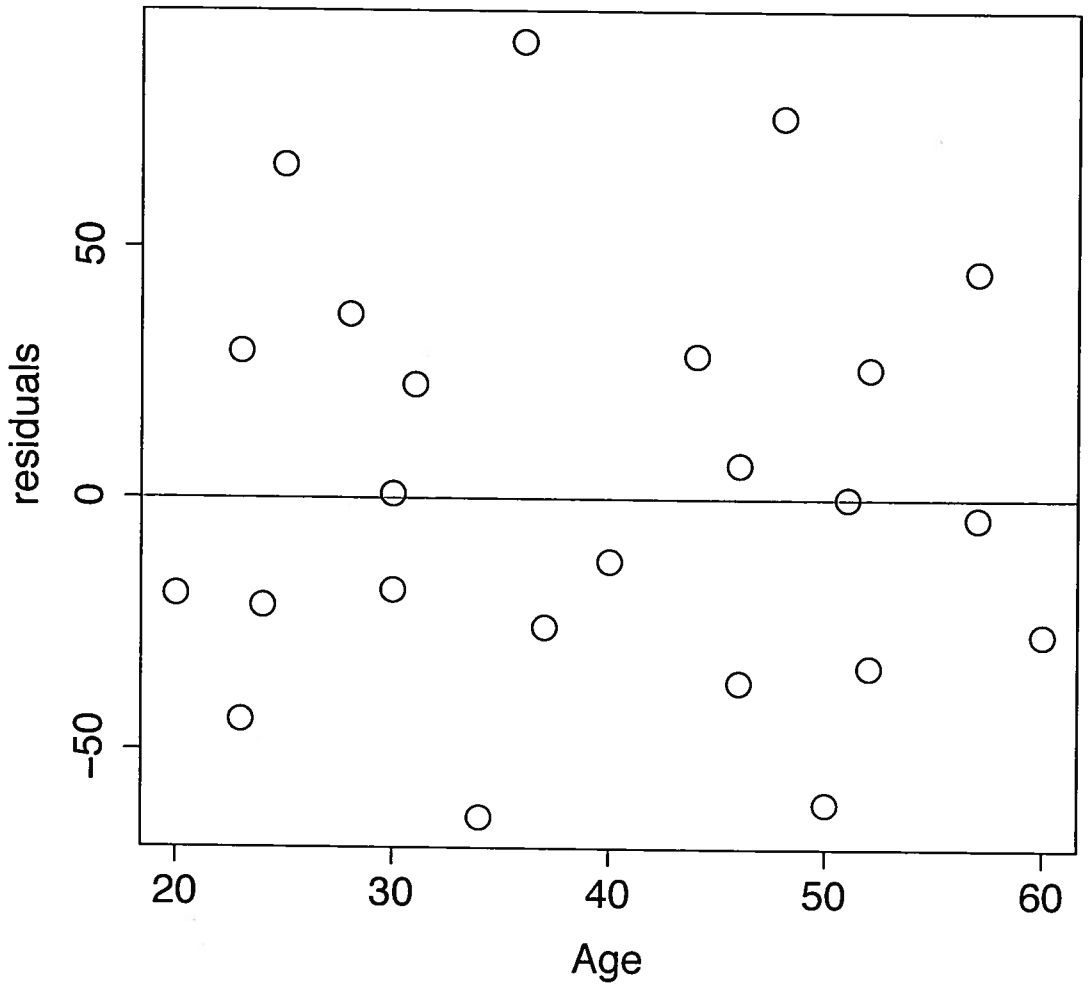
This linear regression line is constructed from data. The slope of this regression model is a statistic and has a sampling distribution, which we can model. We have two models: the linear model, which describes the relationship between blood fat and age as well as a model for the sampling distribution. Whenever we use models we need to check assumptions and conditions.

# Assumptions & Conditions

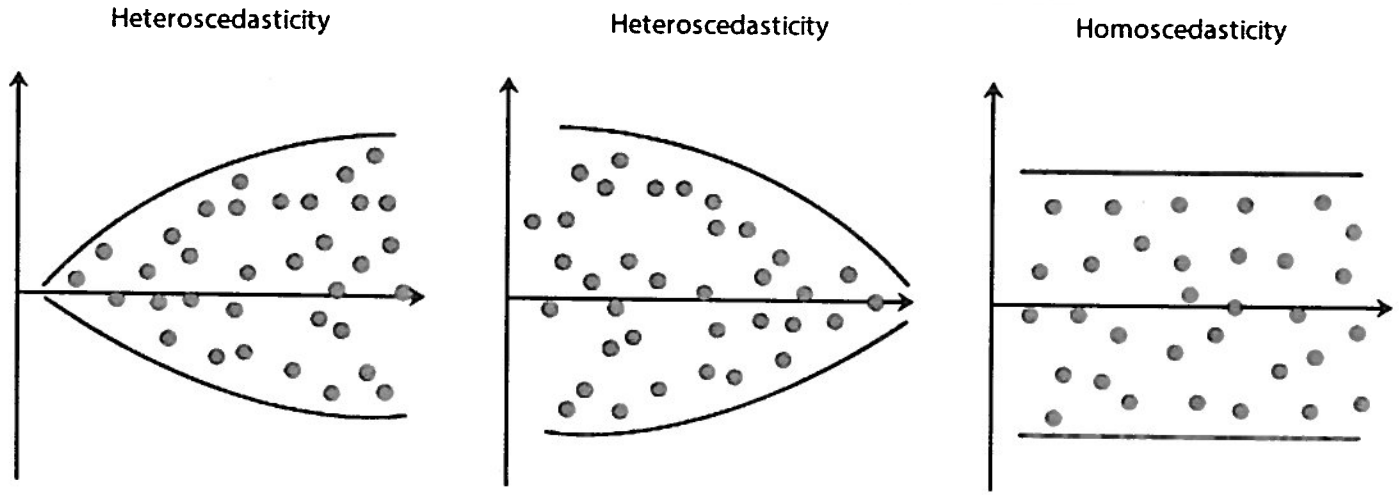
1. Linearity
   - The relationship between $x$ and $y$ must be linear
   - Check this assumption by examining a scatterplot of $x$ and $y$



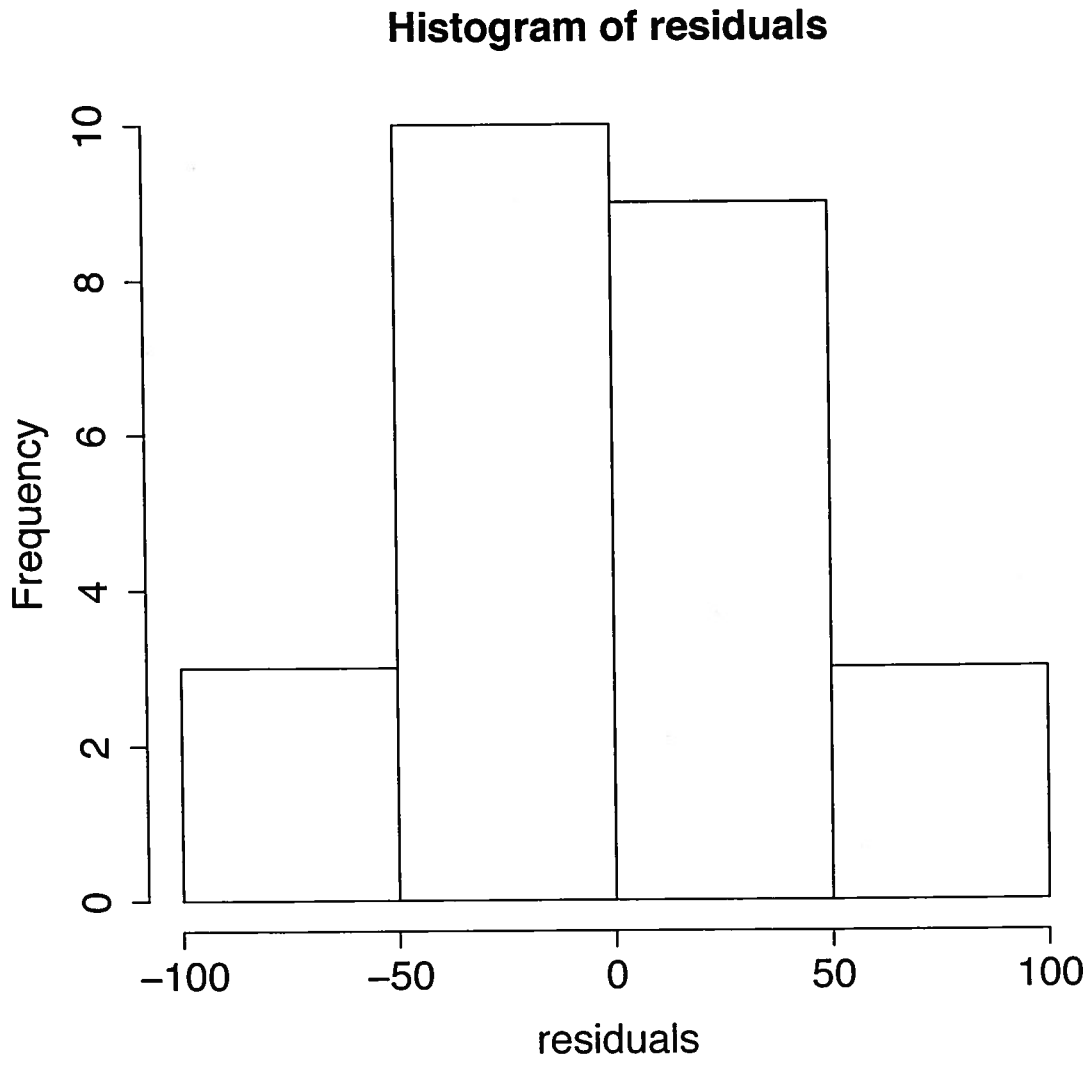
2. Independence of errors
   - $y$ is independent from errors
   - Check by examining a scatterplot of "residuals vs $x$" or "residuals vs. predicted values ($\hat{y}$)".



3. Equal Variances (homoscedasticity)
   - The variance of the residuals should be the same for all values of $x$
   - Check the spread around the line of your scatterplot is nearly constant (check for fan shapes or tendencies of the variation to grow or shrink in one part of the scatterplot). Or look at a "residuals vs. predicted values ($\hat{y}$)" plot; the variance of the residuals should be the same across all values of the x-axis. If the plot shows a pattern (e.g. megaphone shape), then variances are not consistent and this assumption is violated.



4. Normality of errors
   - The residuals must be approximately normal.
   - Examine a histogram of the residuals; it should be nearly normal



# Hypothesis Testing Involving the Slope

estimates

$\hat{y} = b_0 + b_1 x$ sampling distribution.

$b_1 \rightarrow \beta_1$

estimate parameter

1. **Check any necessary assumptions and write the null and alternative hypotheses.**

In simple linear regression, we often wish to test:

Non zero slope. - real effect

horizontal line

$$H_0 : \beta_1 = 0 \quad \text{vs.} \quad H_A : \beta_1 \neq 0$$

(although $H_A : \beta_1 > 0$ or $H_A : \beta_1 < 0$ are also possibilities. In some instances, we test $H_0 : \beta_1 = a$ where $a$ is a number).

A slope of 0 implies the mean value of $y$ for any value of $x$ is the same in which case it means that $x$ is not useful in predicting $y$.

2. **Calculate an appropriate test statistic**

recall: $\frac{\bar{x} - \mu_0}{s/\sqrt{n}}$ ↑ $SE(\bar{x})$

$$t_{obs} = \frac{b_1 - \text{hypothesized value}}{SE(b_1)}$$

$t_{obs} = \frac{b_1 - 0}{SE(b_1)}$

where $SE(b_1) = \frac{s_e}{s_x\sqrt{n-1}}$ ← residual sd

$\frac{b_1 - \beta_1}{SE(b_1)} \sim t_{n-2}$

$$s_e = \sqrt{\frac{\sum(y - \hat{y})^2}{n-2}}$$

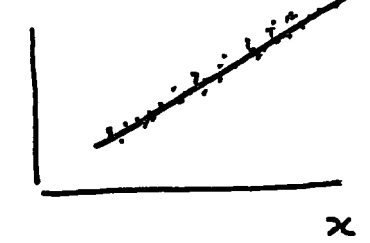
$\hat{y} = b_0 + b_1 x$ (don't know $\beta_0, \beta_1$ must estimate)

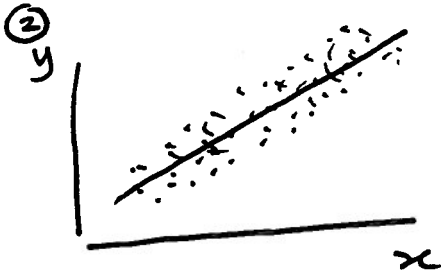
3. **Find the critical region**

Look up a critical value from the $t$-table with $n - 2$ degrees of freedom.

4. **State conclusion**

$s_e$ : scatter around the line

① 

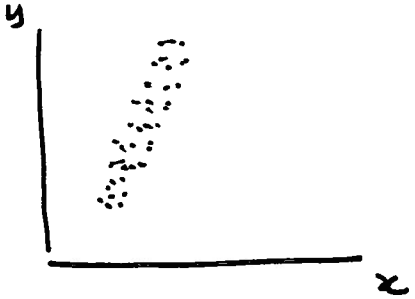

② 


$$SE(b_1) = \frac{s_e}{s_x\sqrt{n-1}}$$

$b_1$

$s_x$ : ① 


higher $s_x$

② 


lower $s_x$

$s_x$ spread of x values

broader x range
→ less varying from sample to sample.

## Confidence interval for the Slope

A $(1-\alpha) \times 100$ confidence interval for $\beta_1$ is:

$$b_1 \pm t^*_{n-2} \times SE(b_1)$$

where the critical value comes from the $t$-distribution with $n-2$ degrees of freedom

## Example 3

A used car dealership in a certain town records the age (years) and price for cars it sold in the last year. Use the output below to answer the following questions.

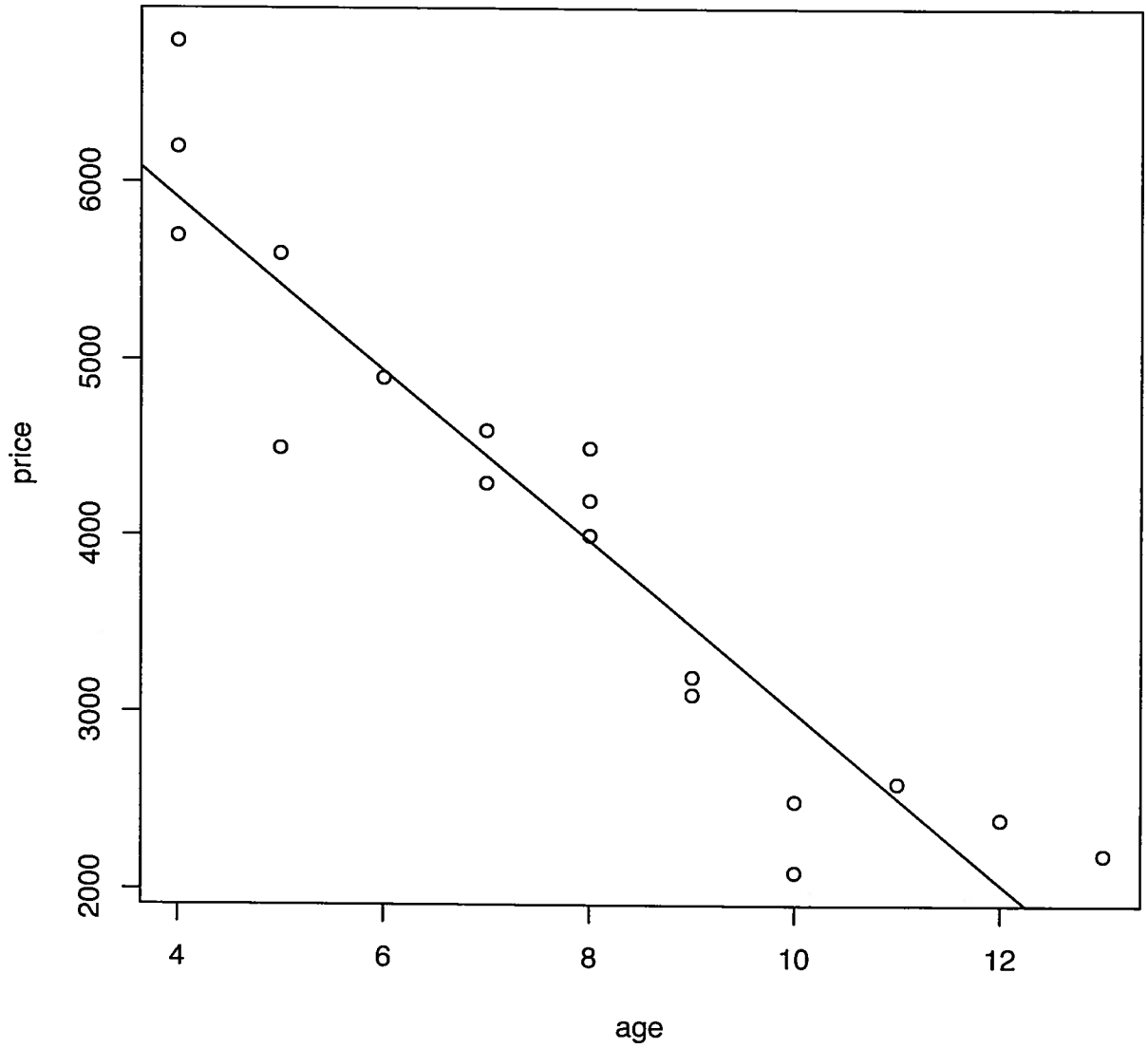
a) Create and interpret a 95% confidence interval for the slope of the regression line.

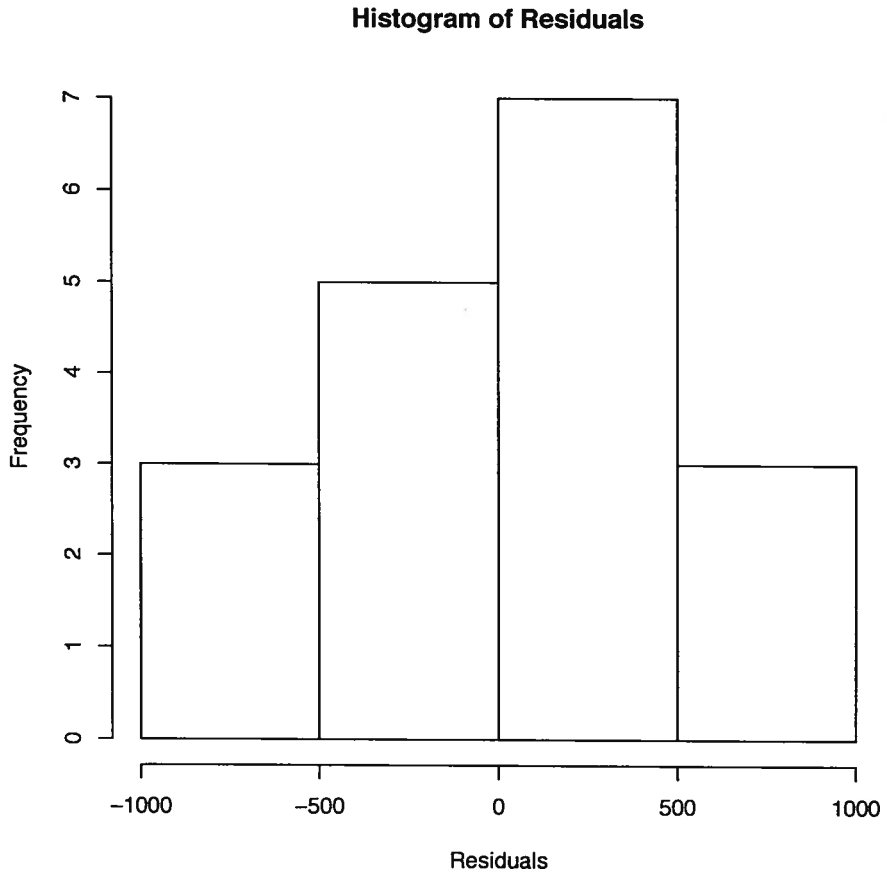
b) Is there evidence of an association between age and price? Test the appropriate hypothesis at a 5% significance level.

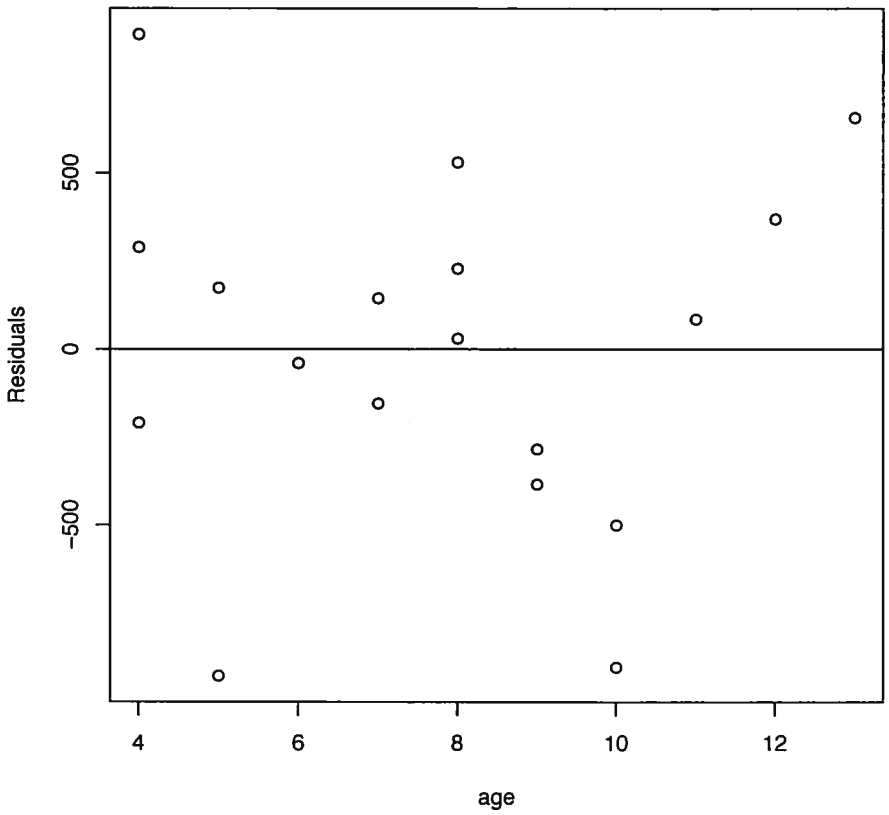
model:

$$\widehat{price} = b_0 + b_1\, age = 7850 - 485\, age.$$

- The price of a car age = 0, model predicts price $7850.
- The model predicts for every additional year the price of car will decrease by $485 on average.


price
2000
3000
4000
5000
6000
4
6
8
10
12
age


Histogram of Residuals
Frequency
0
1
2
3
4
5
6
7
−1000
−500
0
500
1000
Residuals


Residuals
−500
0
500
4
6
8
10
12
age

You should be able to read output from R such as what is shown below.

```
Call:
lm(formula = price ~ age)

Residuals:
   Min     1Q Median     3Q    Max
-925.0 -266.2   57.5  275.0  890.0

Coefficients:
            Estimate Std. Error t value Pr(>|t|)
(Intercept)  7850.00     361.76   21.70 2.71e-13 ***
age          -485.00      43.94  -11.04 6.84e-09 ***
---
Signif. codes:  0 *** 0.001 ** 0.01 * 0.05 . 0.1   1

Residual standard error: 503.1 on 16 degrees of freedom
Multiple R-squared:  0.8839,  Adjusted R-squared:  0.8767
F-statistic: 121.8 on 1 and 16 DF,  p-value: 6.841e-09
```

a) $b_1 = -485$ estimate ± ME

$b_1 \pm t^*_{n-2} SE(b_1)$

$-485 \pm t^*_{16} \; 43.94$ ← R output.

$n = 18$ $\quad -485 \pm 2.120 \times 43.94 = (-578.15, -391.84)$

We are 95% confident the price of used cars decreases an average between $391.84 and $578.15 for each additional year.

b) $H_0: \beta_1 = 0$ (no association btw. age + price)

$H_A: \beta_1 \neq 0$

Check conditions:

① Linearity: scatterplot shows relationship linear.

② Independence: residual plot: no clumping, patterns.

③ Equal variance: residual plot: consistent across.

④ Normality: hist. of residuals nearly normal

Under conditions, sampling dist of slope can be modelled with t-model df = n − 2

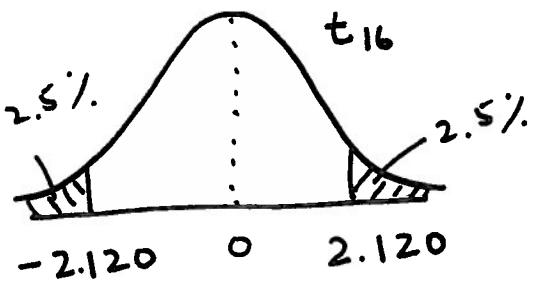
$\alpha = 0.05$

Test statistic:

$$t_{obs} = \frac{b_1 - 0}{SE(b_1)} = \frac{-485}{43.94} = -11.03778$$

↖ agrees with R output

critical region $\alpha = 0.05$




$t_{obs}$ falls in rejection region.

We reject null.

There is strong evidence of an association between age and price.

c) Are price and age negatively linearly related?

$H_0: \beta_1 = 0$ vs. $H_A: \beta_1 < 0$ $\alpha = 0.05$

$t_{obs} = -11.04$


$t_{16,0.05} = -1.746$

$t_{obs}$ falls in rejection region.

Reject $H_0$

Same conclusion above.

Ch. 8, 10, 11

$E(\bar{x}) = \mu$

**one sample z-test.**

$$z_{obs} = \frac{\bar{x} - \mu_0}{\sigma/\sqrt{n}}$$

- $\sigma$ known.
- $\sigma$ unknown + n large.

---

**one sample t-test**

$$t_{obs} = \frac{\bar{x} - \mu_0}{s/\sqrt{n}}$$

$df = n - 1$

- $\sigma$ unknown.
- n small $\rightarrow$ t.

---

**two sample pooled t-test**

$$t_{obs} = \frac{(\bar{x}_1 - \bar{x}_2) - \Delta_0}{s_p\sqrt{\frac{1}{n_1} + \frac{1}{n_2}}}$$

$$s_{pooled}^2 = \frac{(n_1-1)s_1^2 + (n_2-1)s_2^2}{n_1 + n_2 - 2}$$

$df = n_1 + n_2 - 2$

- $\sigma_1^2 = \sigma_2^2 = \sigma^2$
- samples independent.
- both pop. normal or large $n \rightarrow$ CLT.

$H_0: \mu_1 = \mu_2 \rightarrow \mu_1 - \mu_2 = 0$

$H_A: \mu_1 - \mu_2 \; >, <, \neq$

---

**ANOVA F-test.**

$$F_{obs} = \frac{MST}{MSE} = \frac{SST/K-1}{SSE/N-K}$$

$df_1 = K - 1$

$df_2 = N - K$

- 3 or more indep. pop.
- k pop. normal
- $\sigma_1^2 = \ldots = \sigma_k^2$.

t-test slope

$$t = \frac{b_1 - \text{hyp. slope}}{SE(b_1)}$$

$$SE(b_1) = \frac{s_e}{s_x \sqrt{n-1}}$$

$$s_e = \sqrt{\frac{\Sigma (y_i - \hat{y})^2}{n-2}}$$