Topic 6: Correlation and Simple Linear Regression

Optional Reading: Chapter 12

Xiner Zhou

Department of Statistics

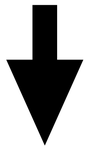
University of California, Davis

One important question in almost all subjects of sciences:

How are two (or more) phenomena related?

- How does aging affect learning rate?
- How does certain medication affect survival rate of a disease?
- How does public insurance provided by government reduce/increase the medical cost and improve/decrease the quality of care?
- How does approximity to technology hub increase individual productivity for tech workers?
- How does reduction in mobility/lockdown decrease COVID-19 transmission?
- ……

i.e. How one variable X is related to another Y?



Correlation

Linear Regression:
is a technique for estimating relationship using data

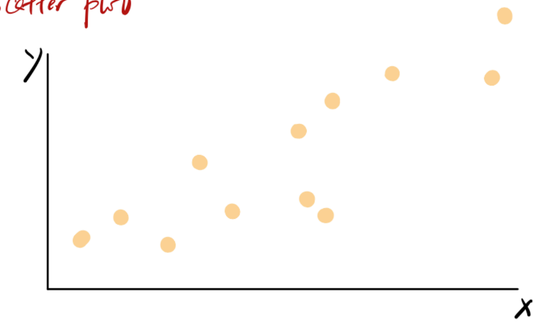
- Relation between two quantities —> Simple Linear Regression
- Relation between many quantities —> multiple linear regression

Correlation

Living Area and Selling Price of 12 Properties

Residence	x (sq. ft.)	y (in thousands)
1	1360	$278.5
2	1940	375.7
3	1750	339.5
4	1550	329.8
5	1790	295.6
6	1750	310.3
7	2230	460.5
8	1600	305.2
9	1450	288.6
10	1870	365.7
11	2210	425.3
12	1480	268.8

Scatter plot



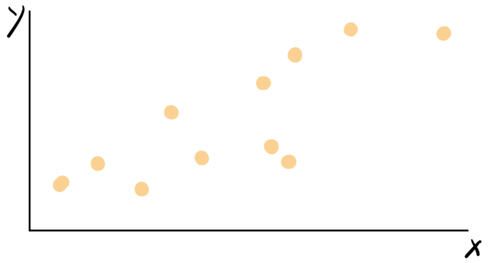
What type of pattern do you see?

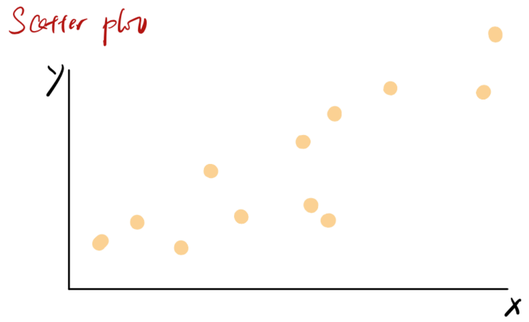
- Is there a constant upward or downward trend that follows a straight-line patter?
- Is there a curved pattern?
- Is there no pattern at all, just a random scattering of points?

Linear relationship:
data points scattered around a straight line, if we were to draw a line through the data cloud

Simple, but cover many important real life scenarios.

Scatter plot





How strong is the pattern?

- Do all of the points follow the pattern exactly?
- Or only weakly?

A single numerical measure that summarizes "How strong is the linear pattern" is called correlation coefficient

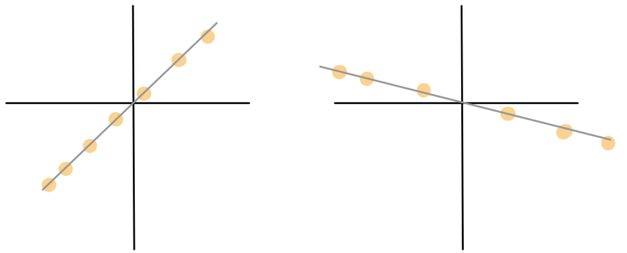
$$r = \frac{S_{XY}}{S_X S_Y}$$

- $S_{XY} = \frac{\sum_{i=1}^{n} (X_i - \bar{X})(Y_i - \bar{Y})}{n-1}$ measures how X and Y move together
  - If positive, it means X and Y tend to move higher or lower together
  - If negative, it means X and Y tend to move in the opposite direction
- $S_X = \sqrt{\frac{\sum_{i=1} (X_i - \bar{X})^2}{n-1}}$ is sample standard deviation for X
- $S_Y = \sqrt{\frac{\sum_{i=1} (Y_i - \bar{Y})^2}{n-1}}$ is sample standard deviation for Y

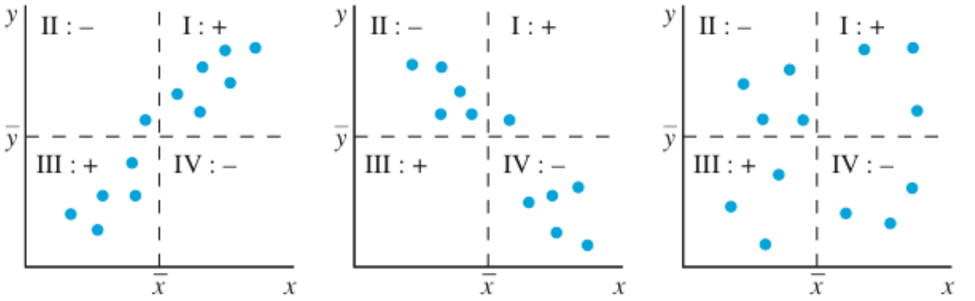
The range of values that correlation coefficient r can take:

$-1 \leq r \leq 1$

- If $0 < r < 1$:
  - X and Y change in the same direction
- If $-1 < r < 0$:
  - X and Y change in the opposite direction
- If $r = 0$:
  - X and Y has no linear relationship
- If $r = \pm 1$:
  - X and Y have perfect linear relationship



$r > 0$ positively correlated $r < 0$ negatively correlated $r = 0$ no correlation



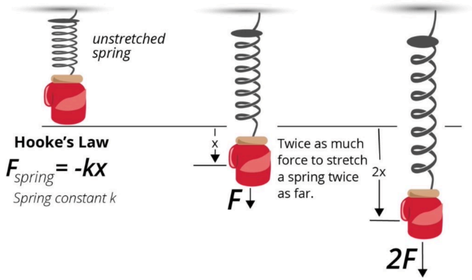
(a) Positive pattern (b) Negative pattern (c) No pattern

- If most of the points are in areas I and III (forming a positive pattern), s_{xy} and r will be positive.
- If most of the points are in areas II and IV (forming a negative pattern), s_{xy} and r will be negative.
- If the points are scattered across all four areas (forming *no* pattern), s_{xy} and r will be close to 0.

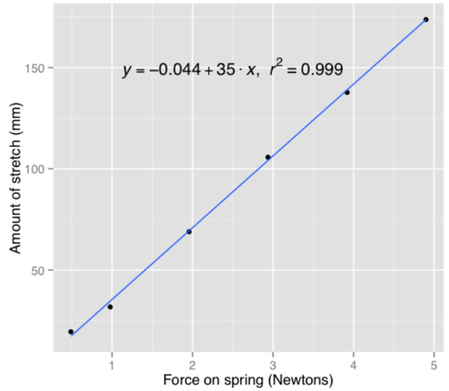
Simple Linear Regression

We are trying to discover or estimate the exact pattern that X and Y follows, using simple linear regression.

Classical mechanics: Hooke's law

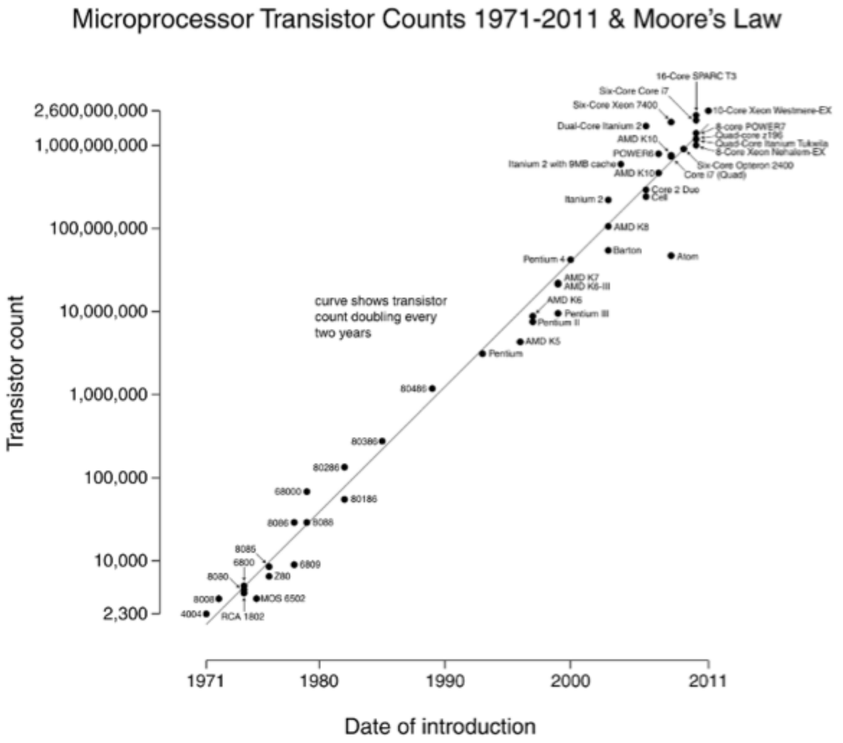


data



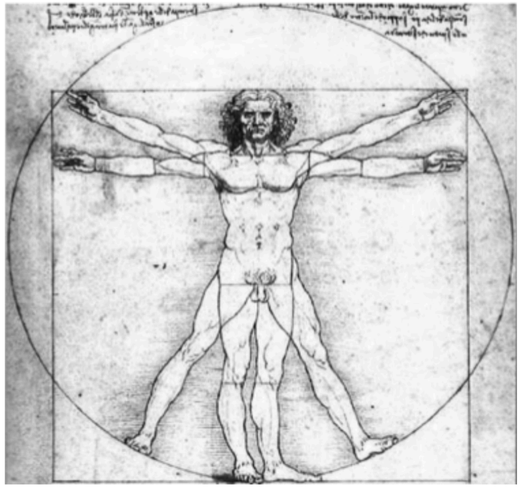
Moore’s law in semiconductor industry:

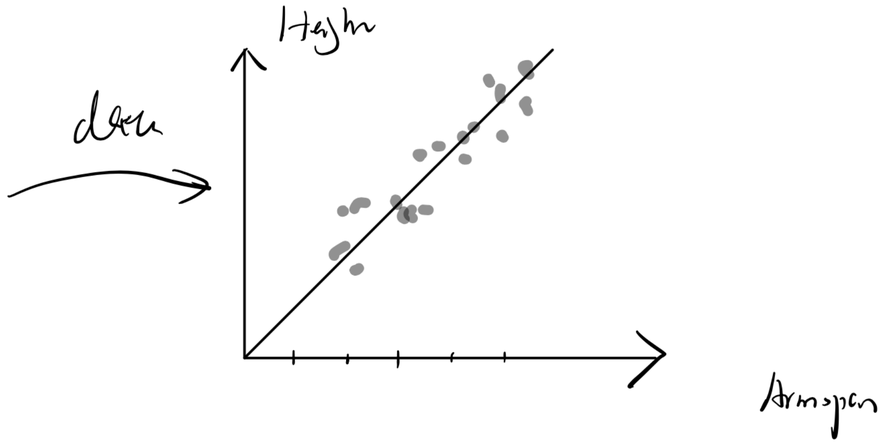
Number of transistors on an integrated circuit doubles roughly every two years.



Leonardo da Vinci: a sketch of a man

Claim: armspan is approximately equal to height





<u>A line fit can discover or verify or help to explain physical phenomena and sciences!</u>

It’s simple, but powerful when used properly.

That’s why linear regression is so popular:

- Medicine
- Scientific papers
- Popular press
- ……..

Suppose we want to model the relationship between r.v.s X and Y,

Usually, Y is some quantity that is likely affected by another quantity X, that is, Y responds to the change of X

Y: dependent variable or response variable
X: independent variable or predictor

We observe paired data points $(x_1, y_1), (x_2, y_2), \ldots (x_n, y_n)$

The simple linear regression model is:

$$Y = \beta_0 + \beta_1 X + \epsilon$$

- A simple linear regression models:
  - response variable Y as a linear function of the predictor X

- ϵ is a noise term
  - In most cases, our data won't fit into a straight line perfectly

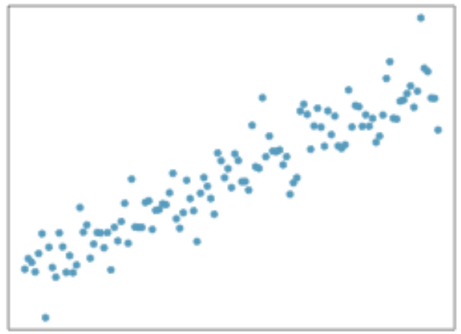
$$\epsilon \sim N(0, \sigma^2)$$

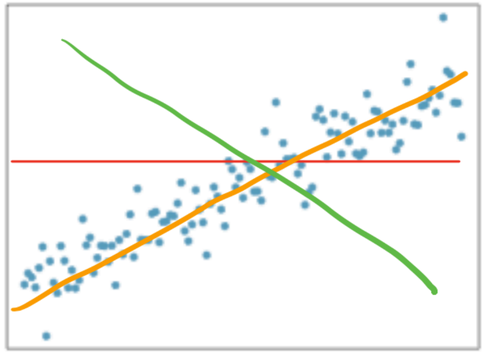
- The main part of the model is the regression line:

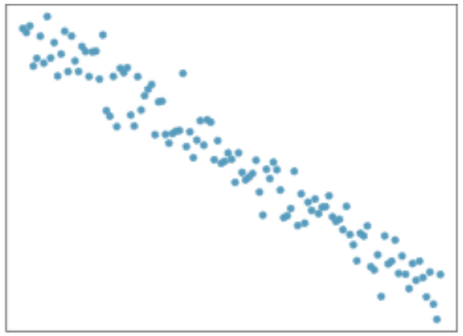
$$\beta_0 + \beta_1 X$$

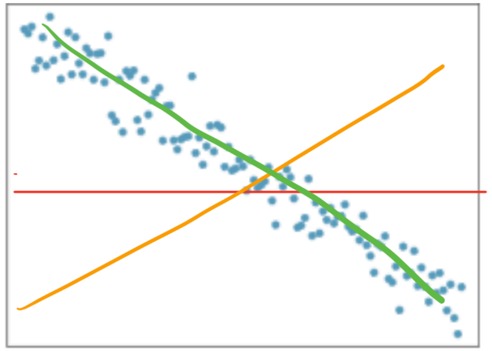
Intercept Slope

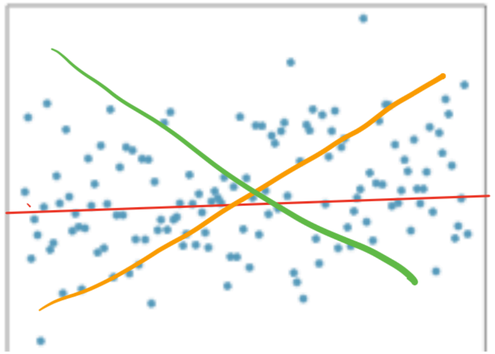
How do we find the best fitting regression line?







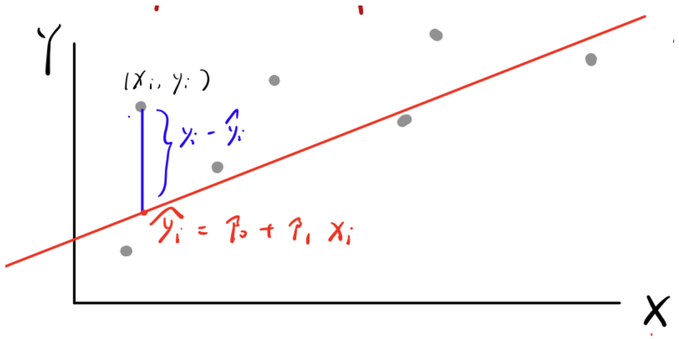




Which line fits the data best?

We need a systematic way to quantify “how good a line fits the data”.

The Least-squares Principle



$\hat{y}_i$: fitted value of the regression model

$y_i - \hat{y}_i$: residual

How to quantify “the best line fits the data’?

Intuition:

since the fitted value represents prediction of regression model, the residual represents the inability or mistakes of the model makes.
So, the natural idea is to find the line with minimum such mistake!

Least-squares principle: to minimize overall residuals as measured by

$$\sum_{i=1}^{n} \left(y_i - \hat{y}_i\right)^2 = \sum_{i=1}^{n} (y_i - \beta_0 - \beta_1 x_i)^2$$

residual sum of squares

$\hat{\beta}_0, \hat{\beta}_1$ are the values that minimize the residual sum of squares

$$\hat{\beta}_1 = \frac{\sum_{i=1}^{n} x_i y_i - \frac{1}{n}(\sum_{1}^{n} x_i)(\sum_{i=1}^{n} y_i)}{\sum_{i}^{n} x_i^2 - \frac{1}{n}(\sum_{i=1}^{n} x_i)^2}$$

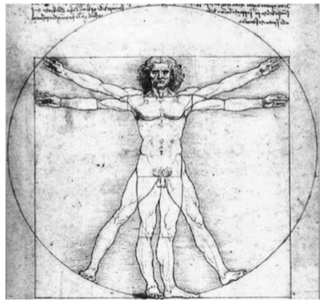
$$\hat{\beta}_0 = \bar{y} - \hat{\beta}_1 \bar{x}$$

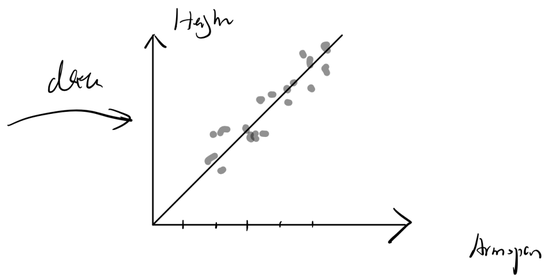
Where $\bar{x} = \frac{1}{n}\sum_{i=1}^{n} x_i$, $\bar{y} = \frac{1}{n}\sum_{i=1}^{n} y_i$

The regression line: $\hat{y} = \hat{\beta}_0 + \hat{\beta}_1 x$

Leonardo da Vinci: a sketch of a man

Claim: armspan is approximately equal to height





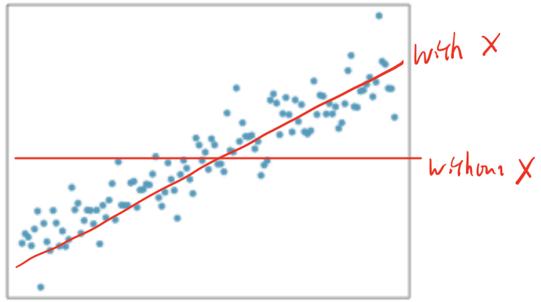
Person	1	2	3	4
Armspan (inches)	68	62.25	65	69.5
Height (inches)	69	62	65	70

Person	5	6	7	8
Armspan (inches)	68	69	62	60.25
Height (inches)	67	67	63	62

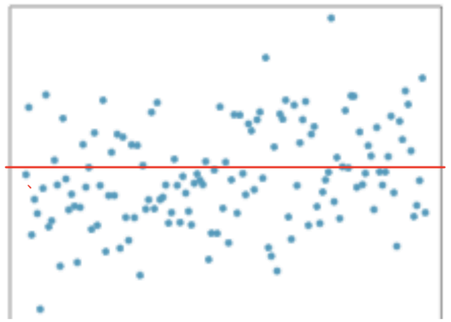
```
> x=c(68,62.25,65,69.5,68,69,62,60.25)
> y=c(69,62,65,70,67,67,63,62)
> lm(y~x)

Call:
lm(formula = y ~ x)

Coefficients:
(Intercept)            x
    12.2214       0.8153
```

Usefulness of the Linear Regression Model

Is the regression line we just fit useful?



Regression using information provided by X substantially better predict Y than without X



Regression using information provided by X predict Y is the same regardless of whether we use or not use X

Predictor or independent variable X is useful in the regression model if:

$$\beta_1 \neq 0$$

i.e. the slope coefficient is not equal to 0

What should we do if we want to prove a claim is true or not?

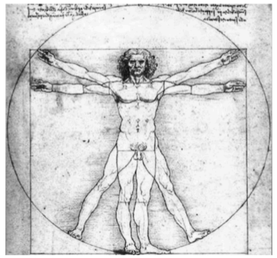
Hypothesis testing:

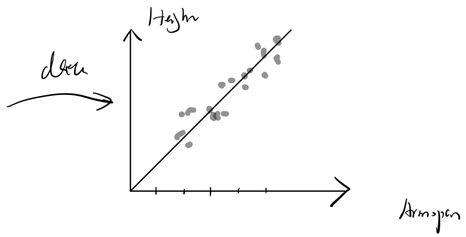
$$H_0 : \beta_1 = 0 \text{ vs } H_a : \beta_1 \neq 0$$

- Same rationale as previous hypothesis testing problems
- Detail not required
- Use t-test, p-value

Leonardo da Vinci: a sketch of a man

Claim: armspan is approximately equal to height





Person	1	2	3	4
Armspan (inches)	68	62.25	65	69.5
Height (inches)	69	62	65	70

Person	5	6	7	8
Armspan (inches)	68	69	62	60.25
Height (inches)	67	67	63	62

To test whether or not armspan is a significant predictor of height, we want to test the hypothesis that

$H_0 : \beta_1 = 0$ vs $H_a : \beta_1 \neq 0$

```
> summary(lm(y~x))

Call:
lm(formula = y ~ x)

Residuals:
     Min       1Q   Median       3Q      Max
-1.47863 -0.74128  0.00564  0.77001  1.33670

Coefficients:
            Estimate Std. Error t value Pr(>|t|)
(Intercept)  12.2214     7.4814   1.634 0.153466
x             0.8153     0.1141   7.148 0.000378 ***
---
Signif. codes:
0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 1.09 on 6 degrees of freedom
Multiple R-squared:  0.8949,    Adjusted R-squared:  0.8774
F-statistic: 51.09 on 1 and 6 DF,  p-value: 0.000378
```

From the t test, since the p-value =0.000378 < 0.05,

We reject H_0 and conclude that a person’s armspan is a significant associated with that person’s height, or a person’s armspan is a significant predictor of person’s height.

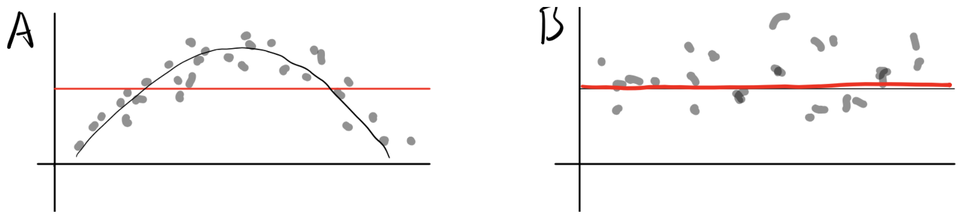
Once the significance or usefulness of the linear regression model is assessed, the model can be used for two purposes:

1. Provide explanation between the relation between X and Y: How does Y change with X?
2. Prediction

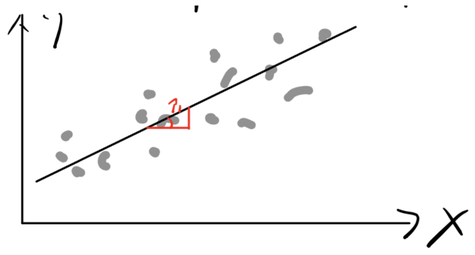
1. How does Y change with X?

Interpretation of the slope coefficient β_1

- If t-test accept $H_0 : \beta_1 = 0$
  - It means no linear relationship between X and Y
  - It's possible that X and Y related inn some other complicated way, but currently we just focus on linear regression



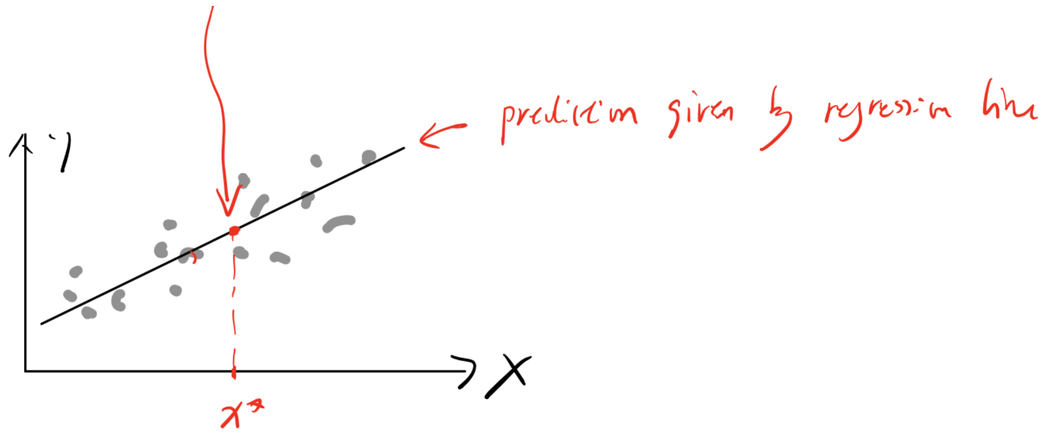
- If t-test accept $H_a : \beta_1 \neq 0$
  - There is a linear relationship between X and Y
  - We can interpret as: for every 1-unit increase in X, there is a β_1-unit increase or decrease in Y



2. Prediction

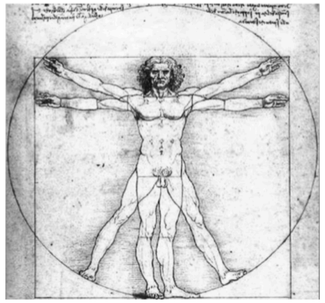
We can use the linear regression model to predict the value of Y at a particular value of $X = x^*$:

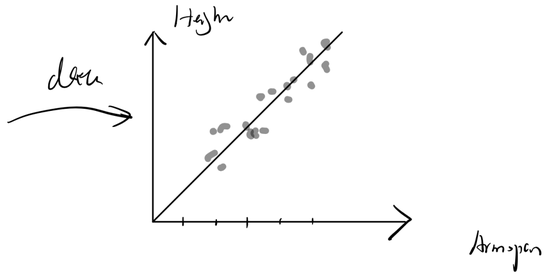
$$\hat{y} = \hat{\beta}_0 + \hat{\beta}_1 x^*$$



Leonardo da Vinci: a sketch of a man

Claim: armspan is approximately equal to height



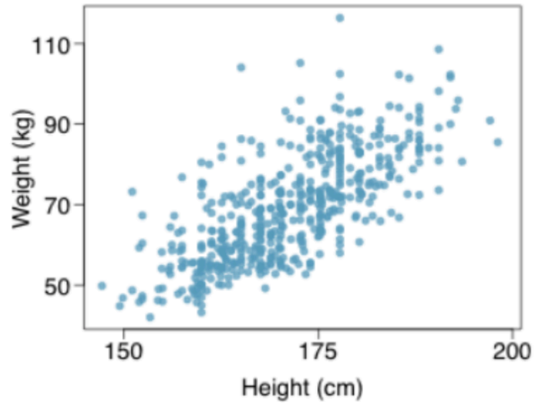


Person	1	2	3	4
Armspan (inches)	68	62.25	65	69.5
Height (inches)	69	62	65	70

Person	5	6	7	8
Armspan (inches)	68	69	62	60.25
Height (inches)	67	67	63	62

```
> summary(lm(y~x))

Call:
lm(formula = y ~ x)

Residuals:
     Min       1Q   Median       3Q      Max
-1.47863 -0.74128  0.00564  0.77001  1.33670

Coefficients:
            Estimate Std. Error t value Pr(>|t|)
(Intercept)  12.2214     7.4814   1.634 0.153466
x             0.8153     0.1141   7.148 0.000378 ***
---
Signif. codes:
0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 1.09 on 6 degrees of freedom
Multiple R-squared:  0.8949,    Adjusted R-squared:  0.8774
F-statistic: 51.09 on 1 and 6 DF,  p-value: 0.000378
```

Body measurements The scatterplot and least squares summary below show the relationship between weight measured in kilograms and height measured in centimeters of 507 physically active individuals.



	Estimate	Std. Error	t value	Pr(> |t|)
(Intercept)	−105.0113	7.5394	−13.93	0.0000
height	1.0176	0.0440	23.13	0.0000

1. Describe the relationship between height and weight.
2. Write the equation of the regression line. Interpret the slope and intercept in context.
3. Do the data provide strong evidence that an increase in height is associated with an increase in weight? State the null and alternative hypotheses, report the p-value, and state your conclusion.
4. What is the predicted weight for a person with height of 180cm?