
# STAT2401
# Analysis of Experiments

## Lecture Week 4

Dr Darfiana Nur

THE UNIVERSITY OF WESTERN AUSTRALIA

# Aims of Lecture Week 4

- **Aim 1 Linear Models**

- **Aim 2 Simple Linear Regression (SLR)** (Sheather Ch 2.1, Moore et al Ch 10)

  2.1 Statistical model

  2.2 Terminology and Assumptions

- **Aim 3 Parameter estimation** (Sheather Ch 2.2, Moore et al Ch 10)

  3.1 The least squares method

  3.2 Interpretation of SLR

- **Aim 4 Statistical Inference: Hypothesis Testing and Confidence Interval** (Sheather Ch 2.2, Moore et al Ch 10)

# Aim 1 Linear models

- A substantial portion of analysis in applied statistics comes under the heading of ***linear models***
- Linear models provide a unified framework for
  - ✓ Fitting linear relationships
  - ✓ $t$-tests, analysis of variance, e.g., *K*–means, other experimental designs
  - ✓ Multivariate analysis
  - ✓ Time-series models
- Key idea is that of an additive *statistical* model, e.g.,

$$y = g(x_0, x_1, x_2, \dots, \beta_0, \beta_1, \beta_2, \dots) + \epsilon$$

along with assumptions about the quantities in the model, in particular, the form of $g(\cdot)$and the error term $\epsilon$

# Example 1 – Fitting a linear relationship

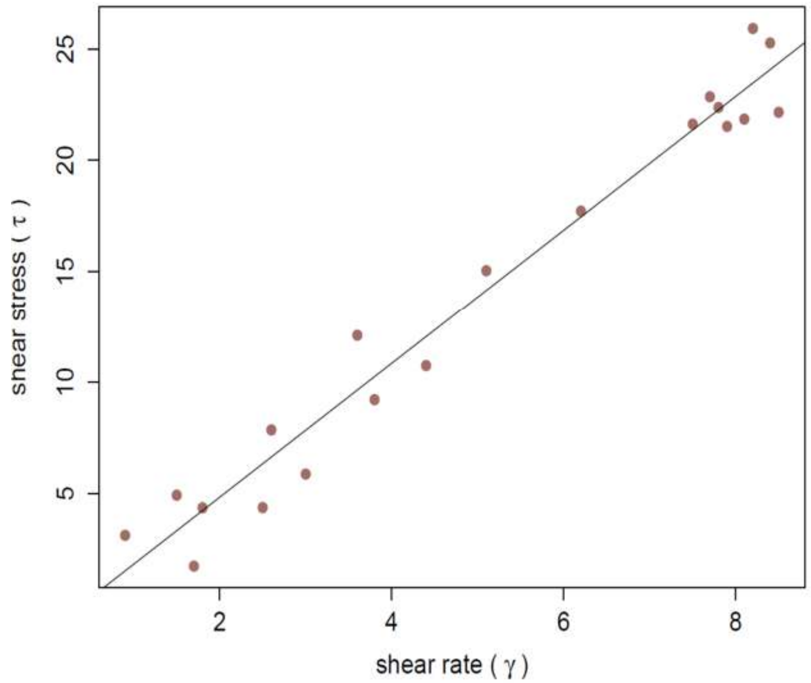
- From theory, the slope of the relationship between shear stress $\tau$ and shear rate $\gamma$ is the viscosity of the fluid $\eta$

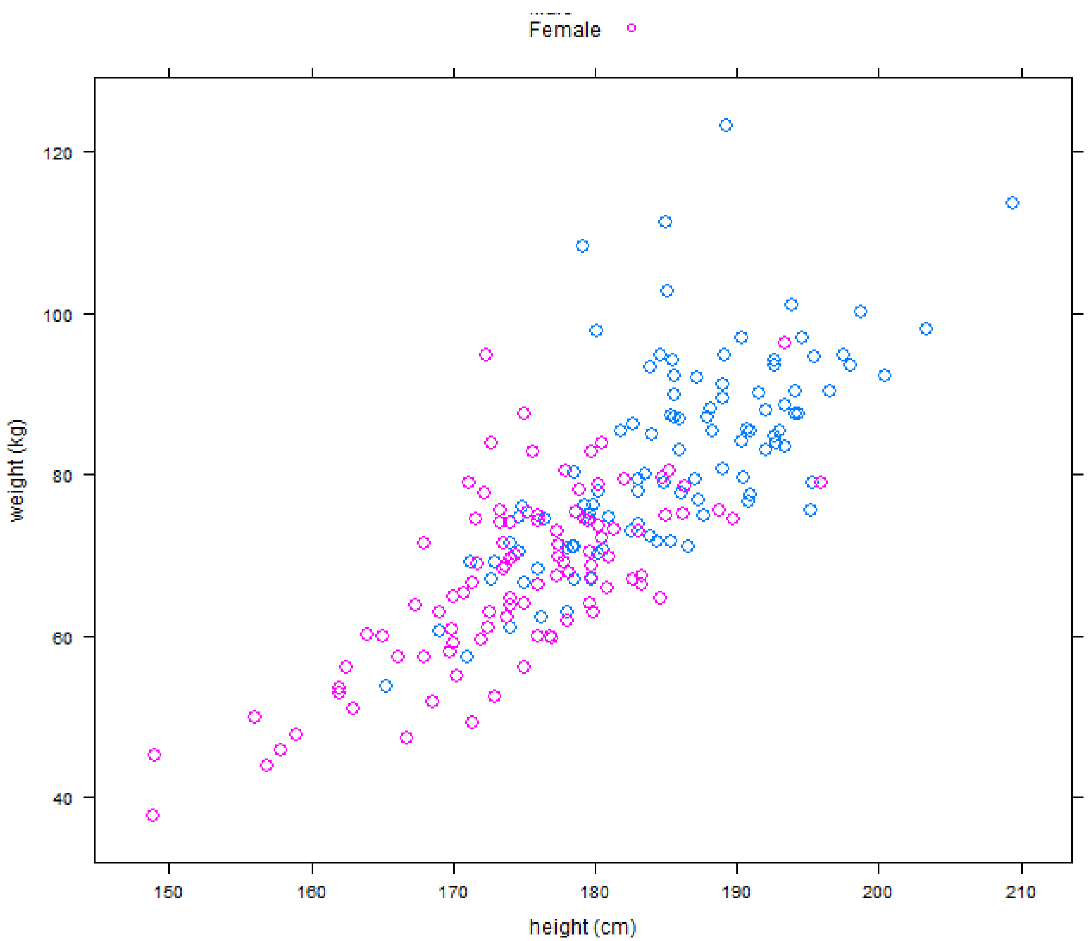
$$\tau = \eta_0 + \eta_1 \gamma$$

- Because of variability and noise – instrumental, raw material – the points are scattered about a straight line
- We postulate a statistical model for the observed data

$$y_i = \beta_0 + \beta_1 x_i + \epsilon_i, \quad \epsilon_i \sim N(0, \sigma^2)$$

- The parameters $\beta_0$ and $\beta_1$ are fixed constants whose values have physical meaning and that we want to estimate






# Example 2 – Fitting a linear relationship

- A linear model that we postulate can be purely empirical
- There is no theoretical relationship that predicts weight from height of AIS athletes, but a linear relationship may be plausible because we know that, on average, taller people weigh more than shorter people
- Could fit a simple linear model of the form

$$y_i = \beta_0 + \beta_1 x_i + \epsilon_i,$$

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

# Why linear models?

We commonly fit linear models because

- In some cases, the underlying relationship is approximately linear
- A simple model might be "good enough" for our purposes
- They might provide a good **approximation** to nonlinear models, e.g., over a narrow region
- It often makes sense to check first if a linear relationship fits; if it doesn't we can fit more complex models

*All models are wrong, but some are useful*

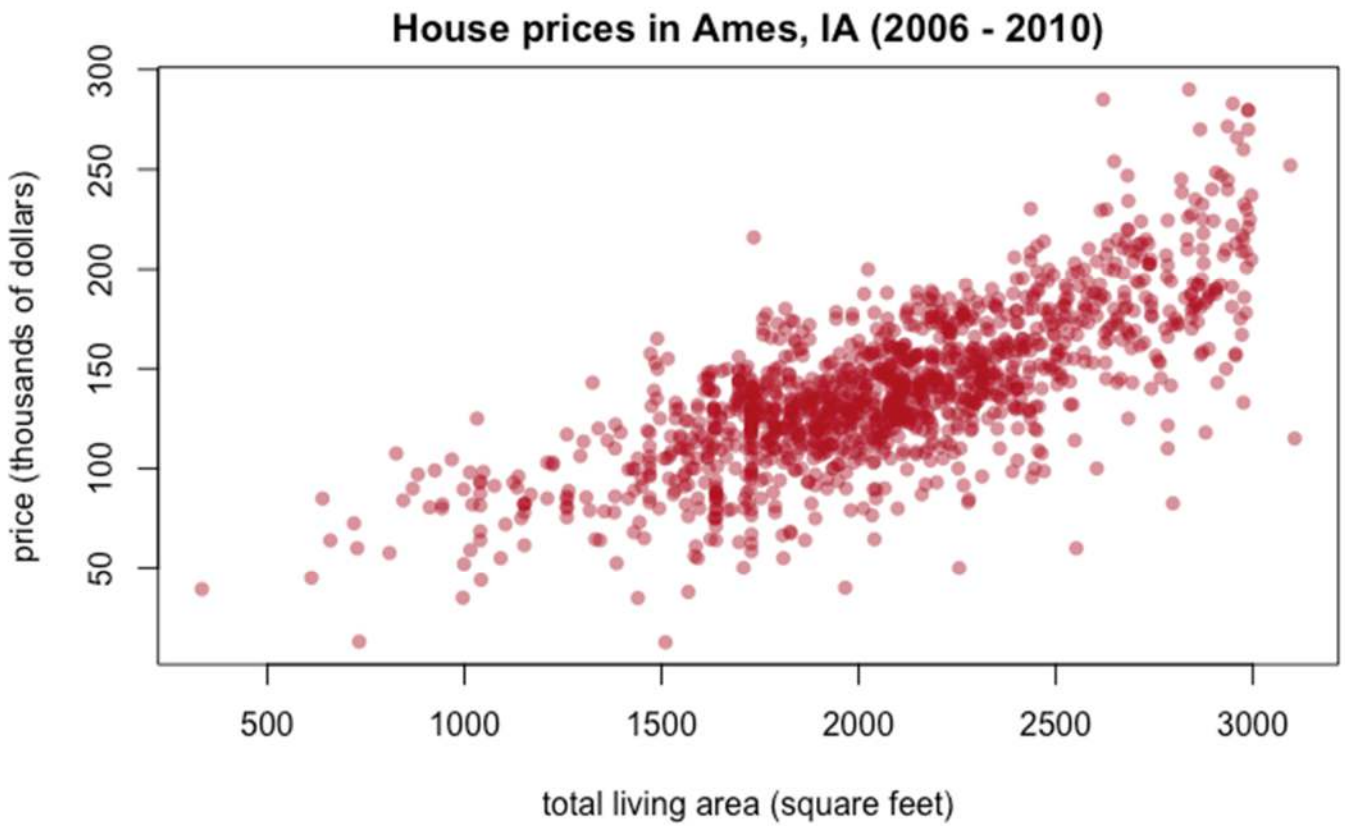
(G.E.P. Box, 1920 – 2013)

# Why linear models?

Linear models provide the basis for learning about extensions such as

- Generalized linear models (GLM)
- Mixed models
- Hierarchical models

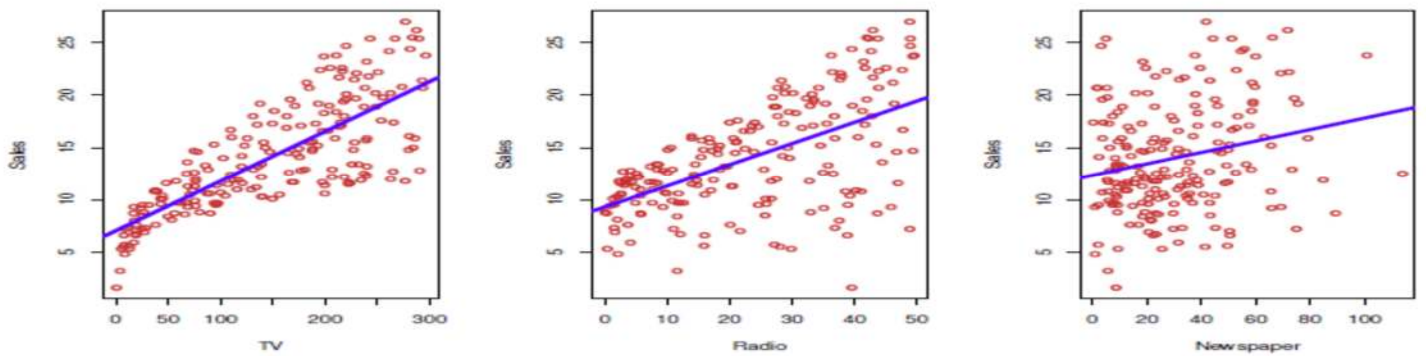
# Linear regression
# Example 3: House Prices data



# Simple Linear Regression

**Example 4.**

- The Advertising data set consists of the sales of that product in 200 different markets, along with advertising budgets for the product in each of those markets for three different media: TV, radio, and newspaper.
- AIM is to investigate the association between advertising and sales of a particular product.




James ET AL (2013, 2023), Ch 2

**FIGURE 2.1.** *The* Advertising *data set. The plot displays* sales, *in thousands of units, as a function of* TV, radio, *and* newspaper *budgets, in thousands of dollars, for 200 different markets. In each plot we show the simple least squares fit of* sales *to that variable, as described in Chapter 3. In other words, each blue line represents a simple model that can be used to predict* sales *using* TV, radio, *and* newspaper, *respectively.*

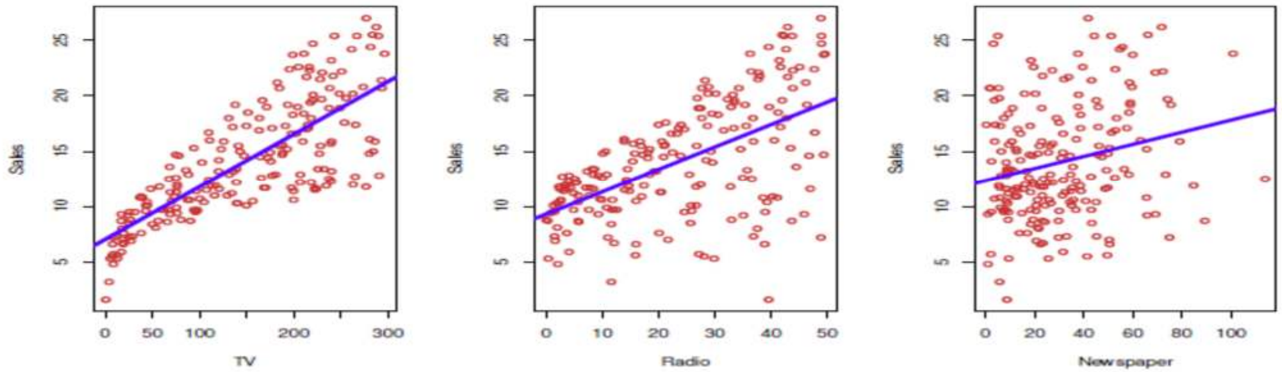
# Example 4 (continued).

Interested in answering questions such as:

– Which media are associated with sales?

– Which media generate the biggest boost in sales?

– How large of an increase in sales is associated with a given increase in TV advertising?




James ET AL (2013, 2021), Ch 2

**FIGURE 2.1.** *The* Advertising *data set. The plot displays* sales, *in thousands of units, as a function of* TV, radio, *and* newspaper *budgets, in thousands of dollars, for 200 different markets. In each plot we show the simple least squares fit of* sales *to that variable, as described in Chapter 3. In other words, each blue line represents a simple model that can be used to predict* sales *using* TV, radio, *and* newspaper, *respectively.*

# Linear regression

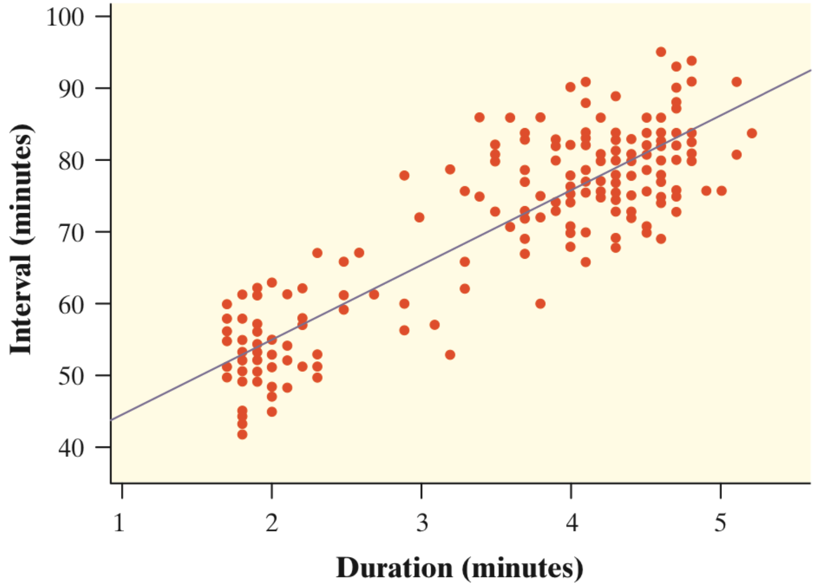
- One of the most common methods in use
  - Fundamental to more complex methods
  - Easy to interpret
  - Efficient to solve
- We make the simple assumption that there is a linear relationship between the **explanatory variable** ($x$, living area) and the **response variable** ($y$, sale price)
- Motivation: given a simple model, we want to be able to predict sale price in the future

# Aim 2 Simple Linear Regression Introduction 1

- When a scatterplot shows a linear relationship between a numerical (quantitative) explanatory variable x and a numerical (quantitative) response variable y, we can use the least-squares line fitted to the data to predict y for a given value of x.
- If the data are a **random sample** from a larger population, we need statistical inference to answer questions like these:
  - ✓ Is there really **a linear relationship between x and y in the population**, or could the pattern we see in the scatterplot plausibly happen just by chance?
  - ✓ What is **the slope (rate of change)** that relates y to x in the population, including a **margin of error for our estimate of the slope**?
  - ✓ If we use the least-squares regression line to predict y for a given value of x, how **accurate** is our prediction (again, with a margin of error)?

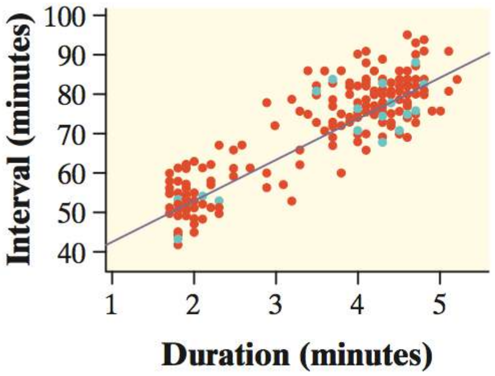
# Simple Linear Regression Introduction 2

**Example 5.** Researchers have collected data on eruptions of the Old Faithful geyser. Here is a scatterplot of the duration and interval of time until the next eruption for all 222 recorded eruptions in a single month. The least-squares regression line for this population of data has been added to the graph. It has slope 10.36 and y intercept 33.97. Regarding all 222 eruptions as the population, this line is the **population regression line** (or true regression line) because it uses all the observations that month.

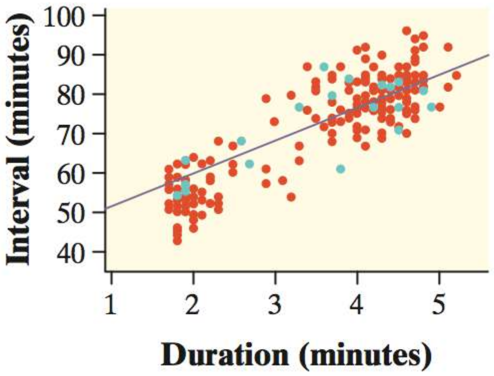



Suppose we take an SRS (Simple Random Sample) of 20 eruptions from the population and calculate the least-squares regression line $\hat{y} = b_0 + b_1 x$ for the sample data. How does the slope of the **sample regression line** (also called the estimated regression line, or LSRL) relate to the slope of the population regression line?

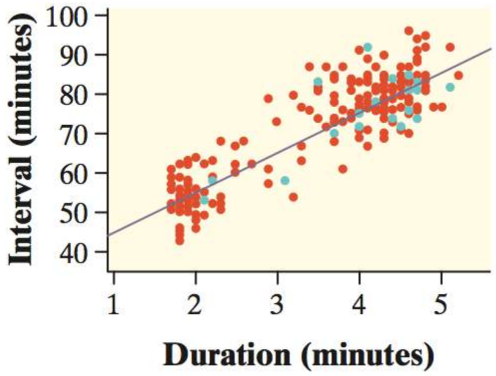
# Simple Linear Regression Introduction 3




Sample 1: $\hat{y} = 32.8 + 10.2x$



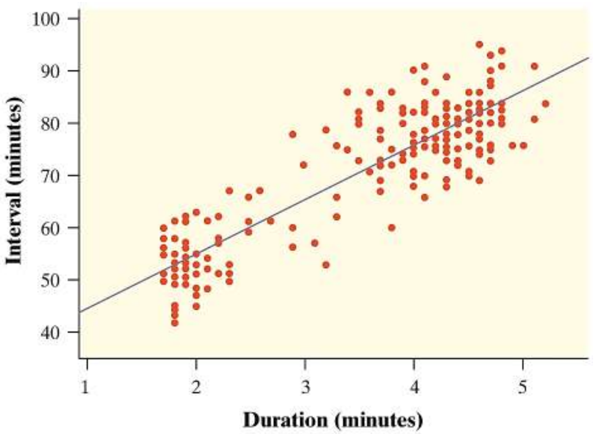

Sample 2: $\hat{y} = 44.0 + 7.7x$




Sample 3: $\hat{y} = 36.0 + 9.5x$

These figures show the results of taking three different SRSs of 20 Old Faithful eruptions in this month.

The green points in each graph are the selected points, and the line is the LSRL for that sample of 20.




Notice that the slopes of the sample regression lines -10.2, 7.7, and 9.5 -vary quite a bit from the slope of the population regression line, 10.36.

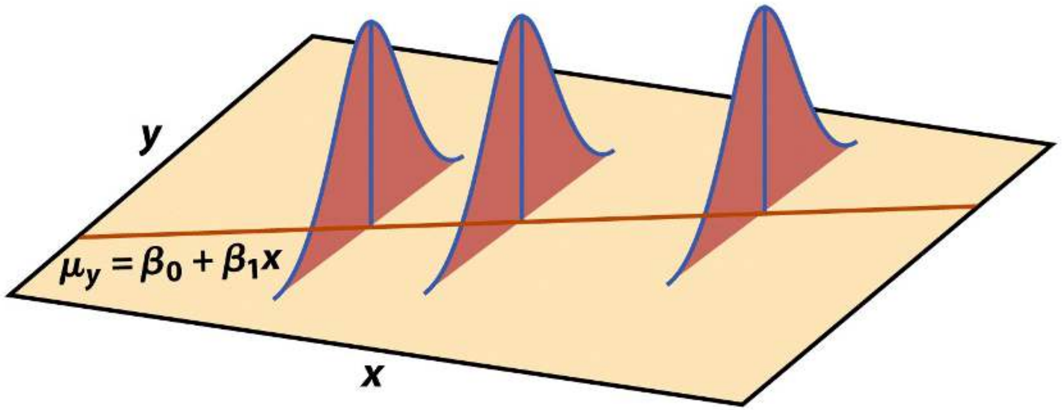
The pattern of variation in the slope *b* is described by its sampling distribution.

# Aim 2.1 Simple Linear Regression (SLR) Model

- In the population, the linear regression equation is $\mu_y = \beta_0 + \beta_1 x$.
- Sample data fits the **simple linear regression model**:

Data = $\boxed{\text{Fit}}$ + $\boxed{\text{Error}}$

$Y_i = (\beta_0 + \beta_1 X_i) + (\varepsilon_i)$




where the $\varepsilon_i$ are **independent** and **Normally** distributed $N(0,\sigma)$.

- Linear regression assumes **equal variance of *y*** ($\sigma$ is the same for all values of x).

# Estimating the Parameters

$$E(Y)=\mu_y = \beta_0 + \beta_1 x$$

- The intercept $\beta_0$, the slope $\beta_1$, and the standard deviation $\sigma$ of $y$ are the unknown parameters of the regression model. We rely on the random sample data to provide unbiased estimates of these parameters.
- The value of $\hat{y}$ from the least-squares regression line is really a prediction of the mean value of $y$ ($\mu_y$) for a given value of x.
- The least-squares regression line ($\hat{y} = b_0 + b_1 x$) obtained from sample data is the best estimate of the true population regression line ($\mu_y = \beta_0 + \beta_1 x$).

$\hat{y}$ is an unbiased estimate for mean response $\mu_y$

$b_0$ is an unbiased estimate for intercept $\beta_0$

$b_1$ is an unbiased estimate for slope $\beta_1$

$$\widehat{\beta_0} = b_0$$

$$\widehat{\beta_1} = b_1$$

# Conditions for Regression Inference 1

- The slope and intercept of the least-squares line are *statistics*. That is, we calculate them from the sample data. These statistics would take somewhat different values if we repeated the data production process. To do inference, think of $b_0$ and $b_1$ as estimates of unknown parameters $\beta_0$ and $\beta_1$ that describe the population of interest.

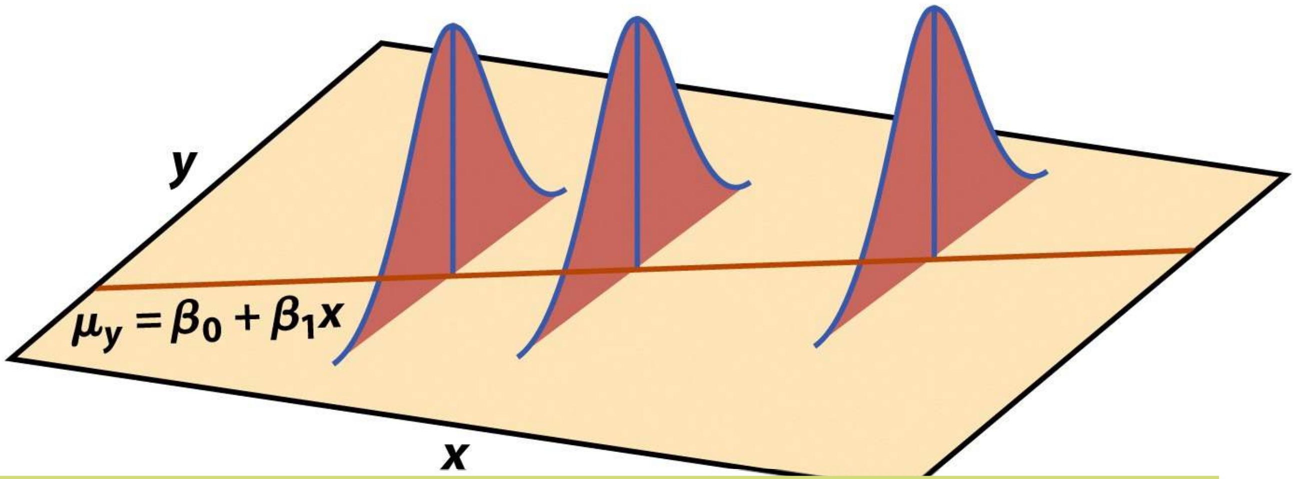
**Conditions for Regression Inference**

We have *n* observations on an explanatory variable *x* and a response variable *y*. Our goal is to study or predict the behavior of *y* for given values of *x*.

- For any fixed value of *x*, the response *y* varies according to a **Normal distribution.** Repeated responses *y* are **independent** of each other.
- The mean response $\mu_y$ has a **straight-line relationship** with *x* given by a population regression line $\mu_y = \beta_0 + \beta_1 x$.
- The slope and intercept are unknown parameters.
- **The standard deviation of *y* (call it $\sigma$) is the same** for all values of *x*. The value of $\sigma$ is unknown.

# Conditions for Regression Inference 2

- The figure below shows the regression model when the conditions are met. The line in the figure is the population regression line $\mu_y = \beta_0 + \beta_1 x$.




The Normal curves show how y will vary when x is held fixed at different values.
All the curves have the same standard deviation $\sigma$, so the variability of *y* is the same for all values of x.

For each possible value of the explanatory variable x, the mean of the responses $\mu(y \mid x)$ moves along this line.

The value of $\sigma$ determines whether the points fall close to the population regression line (small $\sigma$) or are widely scattered (large $\sigma$).

# Aim 2.2 Simple linear regression (SLR) – terminology

- In the AIS data (Example 2), we have pairs of observations
$$(x_1, y_1), (x_2, y_2), \dots, (x_n, y_n)$$
- The $x_i$ are the values of the **explanatory** or **predictor** variable $X$ (height), and the $y_i$ denote the values of the **response** variable $Y$ (weight)
- Distinction between $X$ and $Y$ is important because we may wish to
  - predict $Y$ from $X$
  - explain the variability in $Y$ by $X$
  - control $Y$ using $X$

# SLR – Model and assumptions

- To complete the specification of the model, we assume
  1. $E(\epsilon_i) = 0$, for all $i$
  2. $\text{var}(\epsilon_i) = \sigma^2$, for all $i$
  3. $\epsilon_i$ and $\epsilon_j$ are independent for all $i \neq j$
  4. $\epsilon_i \sim N(0, \sigma^2)$ if we wish to make inferences about the regression model

- The assumptions imply that

$$E(Y \mid X = x) = \beta_0 + \beta_1 x \text{ and}$$
$$\text{var}(Y \mid X = x) = \sigma^2$$

and hence that if we have repeated observations at different values of $x$, the scatter about the true line will be Normally distributed with constant variance $\sigma^2$

# AIM 3.1 The least-squares (regression) line

- A least-squares or regression line is a straight line that describes how a response variable *y* changes as an explanatory variable x changes.

- We often use a regression line to predict the value of *y* for a given value of x.

- In regression, the distinction between explanatory and response variables is important.

# Which line?

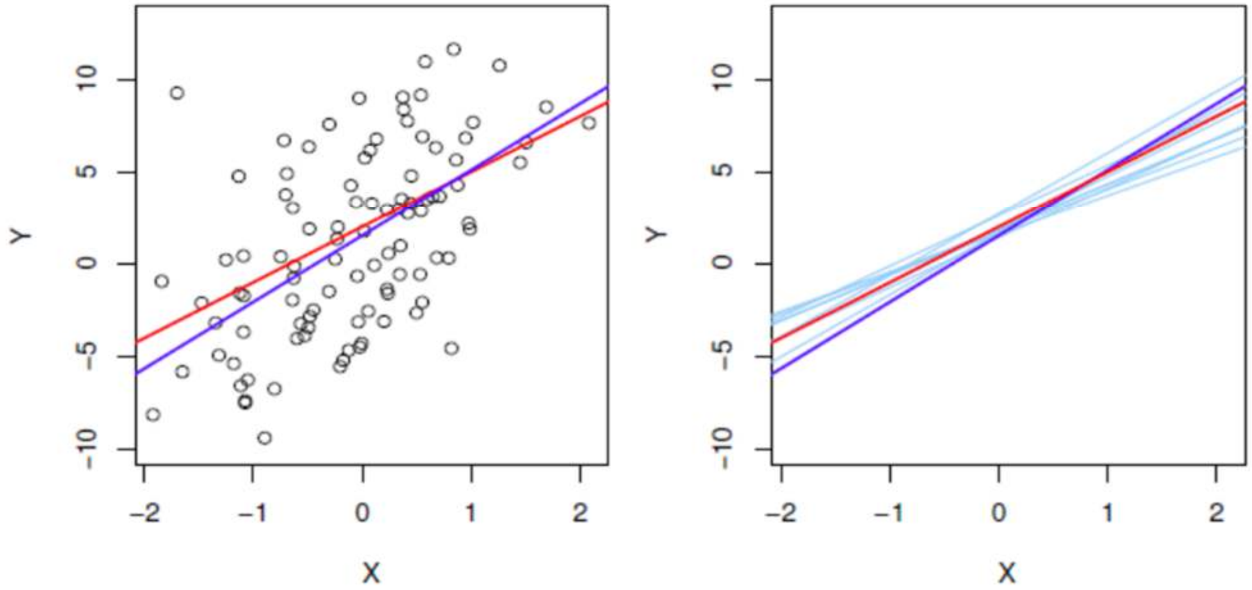



**FIGURE 3.3.** *A simulated data set.* Left: *The red line represents the true relationship, $f(X) = 2 + 3X$, which is known as the population regression line. The blue line is the least squares line; it is the least squares estimate for $f(X)$ based on the observed data, shown in black.* Right: *The population regression line is again shown in red, and the least squares line in dark blue. In light blue, ten least squares lines are shown, each computed on the basis of a separate random set of observations. Each least squares line is different, but on average, the least squares lines are quite close to the population regression line.*

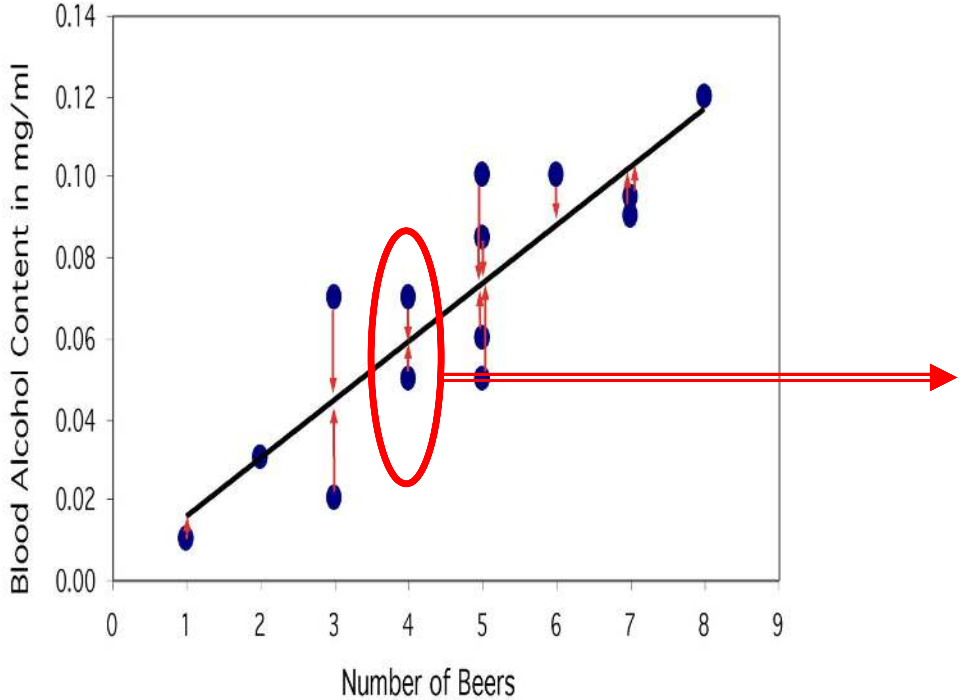
James ET AL (2013, 2021), Ch 3 page 64

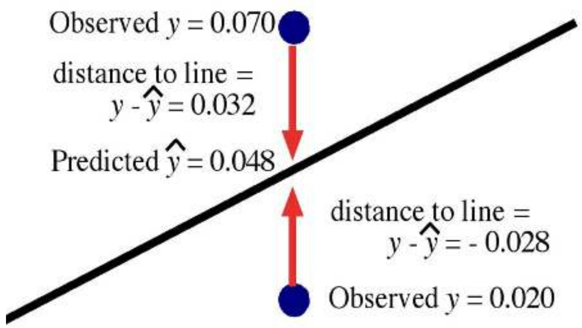
# The least squares method

- When a line is drawn on a scatterplot, we wish to have the vertical distances of the observations from the line to be as small as possible.
- The method that achieves this is known as the **least squares method** and the line that is obtained using this method is known as the **least squares regression line.**

# The least-squares (regression) line

The least-squares regression line is the unique line such that the sum of the squared vertical distances $(y - \hat{y})$ between the data points/observed y and the predicted $\hat{y}$ on the line is as small as possible.




Distances between the points and line are squared so all are positive values. This is done so that distances can be properly added (Pythagoras).

# Least Squares estimation

- We wish to choose the straight line that minimises Sum of Squares of Error (SSE) $SSE = \sum (Y_i - \beta_0 - \beta_1 X_i)^2$
  - To do this we must set the 1st partial derivatives of this formula to 0.

$$\frac{\partial SSE}{\partial \beta_0} = -2\sum (Y_i - \beta_0 - \beta_1 X_i) = 0$$

**Data = Fit + Error**

**$Y_i$ = $(\beta_0 + \beta_1 X_i)$ + $(\varepsilon_i)$**

$$\frac{\partial SSE}{\partial \beta_1} = -2\sum X_i (Y_i - \beta_0 - \beta_1 X_i) = 0$$

After re-arranging some terms we have

$$\sum Y_i = n\beta_0 + \beta_1 \sum X_i$$

$$\sum X_i Y_i = \beta_0 \sum X_i + \beta_1 \sum X_i^2$$

These are called the *normal equations* and must be solved to provide the estimates $\hat{\beta}_0, \hat{\beta}_1$

$\widehat{\beta_0} = b_0$

$\widehat{\beta_1} = b_1$

# SLR – Least squares estimation

- Rearranging the equations on the previous slide yields

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

$$\beta_0 \sum_{i=1}^{n} x_i + \ \beta_1 \sum_{i=1}^{n} x_i^2 = \sum_{i=1}^{n} x_i y_i$$

- These equations are known as the **normal equations**, and solving them yields the least squares estimates of the intercept and slope

# Least Squares estimation

- We can easily solve these two equations given some data points $Y$ and $X$.

$$\hat{\beta}_0 = \overline{Y} - \hat{\beta}_1 \overline{X}$$

$$\hat{\beta}_1 = \frac{\sum (X_i - \overline{X})(Y_i - \overline{Y})}{\sum (X_i - \overline{X})^2} \approx \frac{Cov(X,Y)}{Var(X)} \qquad \frac{S_{xy}}{S_{xx}}$$

- It is straightforward to show, using the second order partial derivatives that this point is a minimum for the SSE

$$\widehat{\beta_0} = b_0$$

- Luckily, R calculates these for us!

$$\widehat{\beta_1} = b_1$$

# Properties of least squares estimators

AUSTRALIA

- The least squares estimators are **unbiased**

$$\widehat{\beta_0} = b_0$$

$$E(\hat{\beta}_0) = \beta_0; E(\hat{\beta}_1) = \beta_1$$

$$\widehat{\beta_1} = b_1$$

- What does this mean?

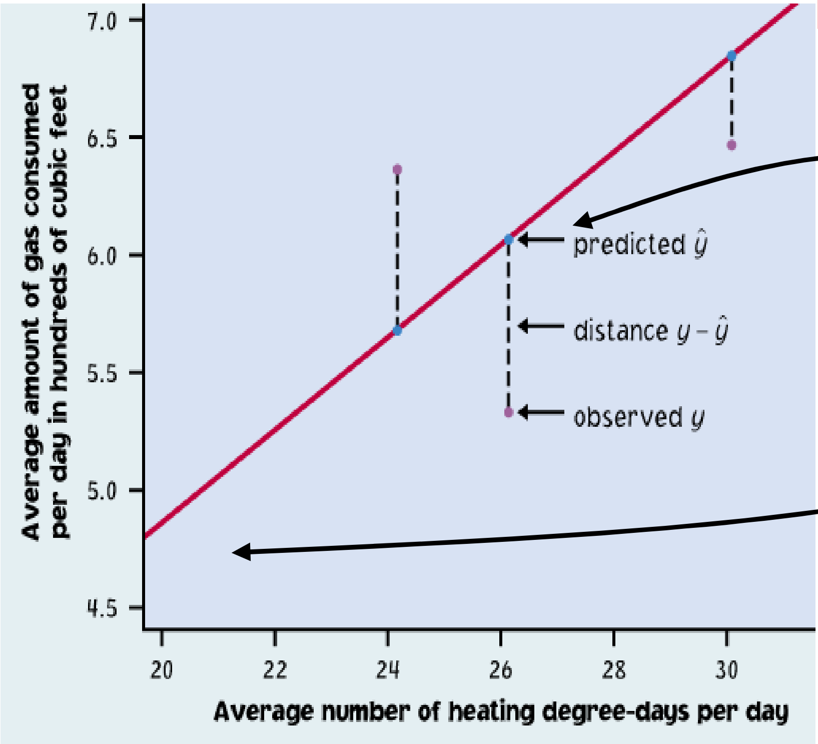
$\hat{\beta}_0, \hat{\beta}_1$ are random variables, they are subject to variation in different samples

- If you take lots of samples and then take the average of the estimates of $\hat{\beta}_0, \hat{\beta}_1$ these will be equal to the true population values $\beta_0, \beta_1$

# Properties

The least-squares regression line can be shown to have this equation:

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



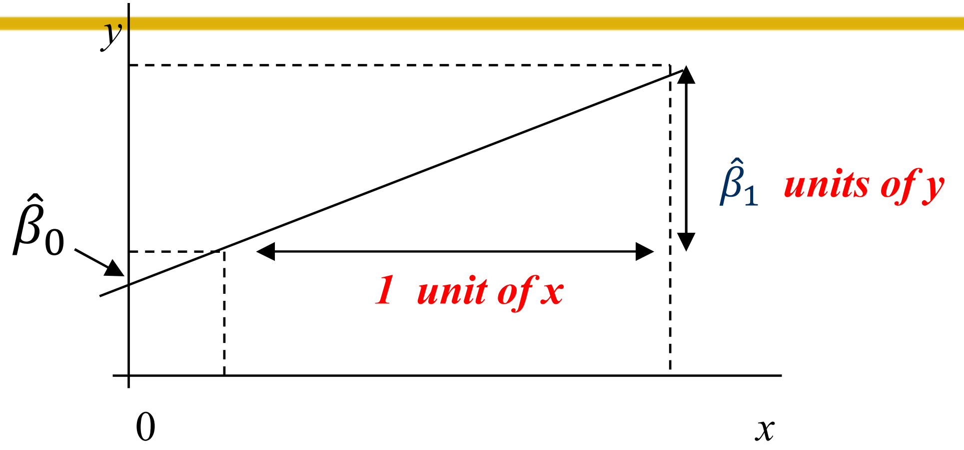

$\hat{y}$ is the predicted $y$ value (y hat)

$\hat{\beta}_1$ is the **slope**

$\hat{\beta}_0$ is the ***y*-intercept**

# Aim 3.1 The interpretation

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




$\hat{\beta}_0$ ➡ = intercept

$\hat{y}$ = value at x = 0

Interpretable only if x = 0 is a value of practical interest

$\hat{\beta}_1$ ➡ = slope

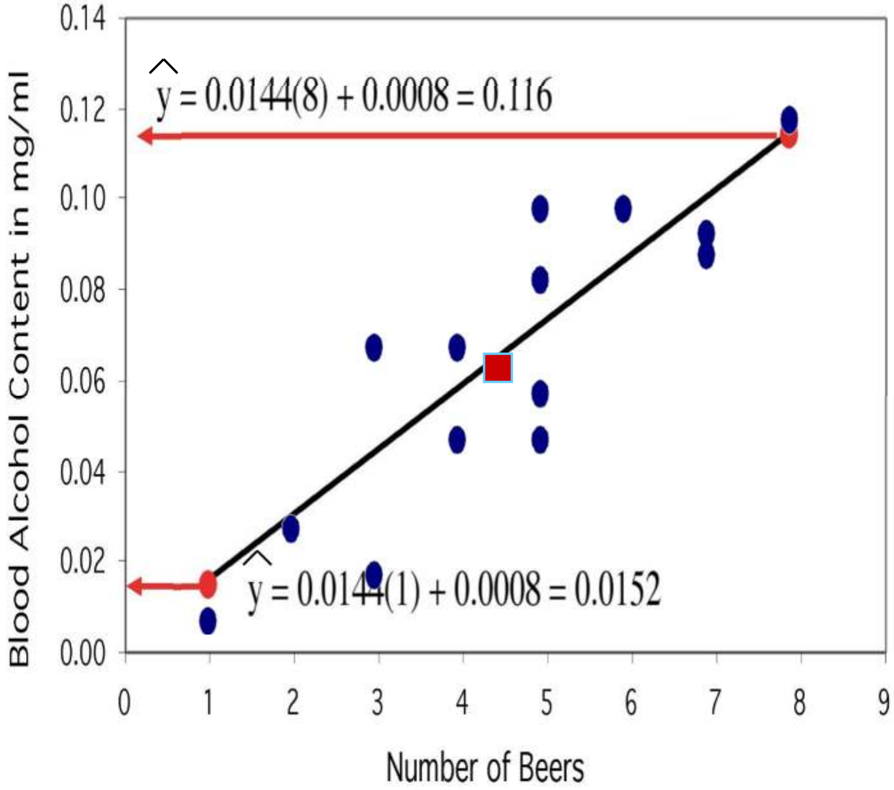
$\hat{y}$ = change in for every 1-unit increase in x

Always interpretable

The equation completely describes the regression line.

To plot the regression line you only need to plug two *x* values into the equation, get *y*, and draw the line that goes through those points.

*Hint: The regression line always passes through the mean of* x *and* y.




$$\hat{y} = 0.0008 + 0.0144x$$

The points you use for drawing the regression line are derived from the equation.

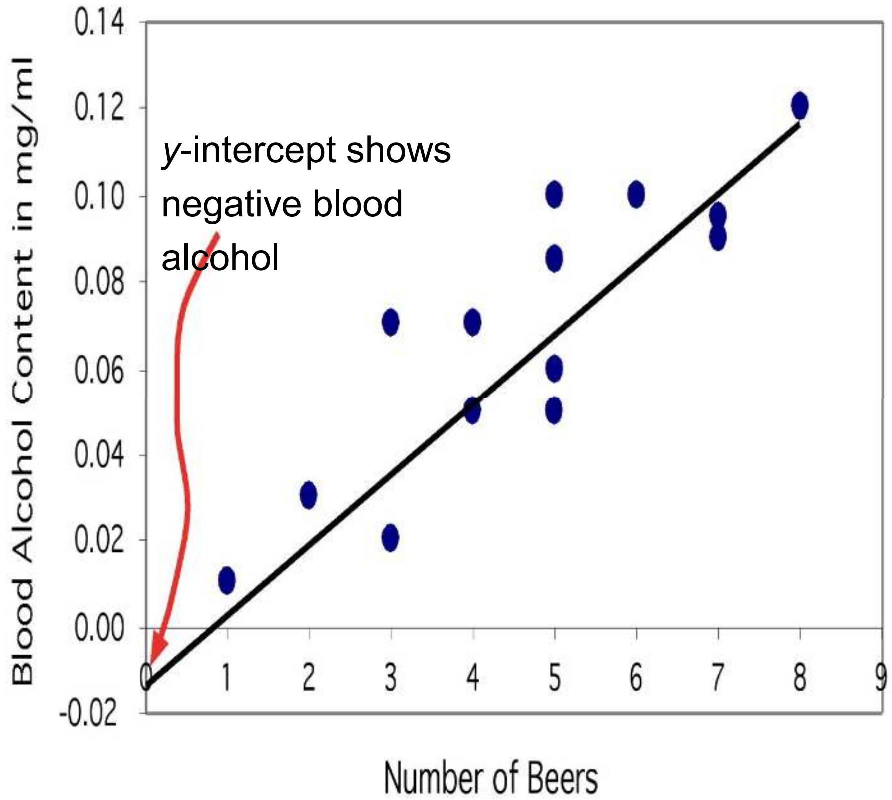
They are NOT points from your sample data (except by pure coincidence).

# The *y* intercept

Sometimes the *y*-intercept is not biologically possible. Here we have negative blood alcohol content, which makes no sense…

But the negative value is appropriate for the equation of the regression line.

There is a lot of scatter in the data, and the line is just an estimate.

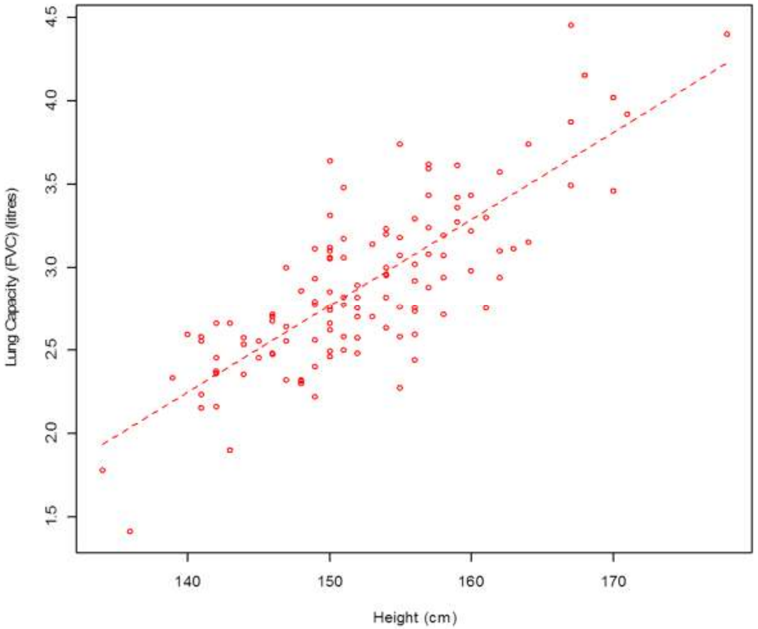


# Example 7: The Boys' FVC data

- In a study on lung function, lung capacity (FVC) in litres and height (cms) were measured on 127 twelve year old boys
- The purpose of the study was to define the range of "normal" FVC's for boys of that age.
- It is essential to recognize that height is an important determinant of FVC. Eg, normal FVC for a 170cm boy will be higher than for a 140cm boy.
- Our purpose is therefore to quantify this relationship.
- In this application, height is the predictor x and FVC is the response y.

# Example 7: The Boys' FVC data (continued)

For many datasets, the linear relationship can be summarised by finding the "line of best fit" or "regression line" or "least-squares line"




$$\hat{y} = -5.065 + 0.052x$$

- **Slope interpretation (B=0.052):** The lung capacity FVC (y) in litres is predicted to increase by 0.052 (slope) for every one cm increase in height.
- **Intercept interpretation (A=-5.065):** Not interpretable as Height (x)=0 is not of practical interest in this scenario.

# Software output – R for the FVC data

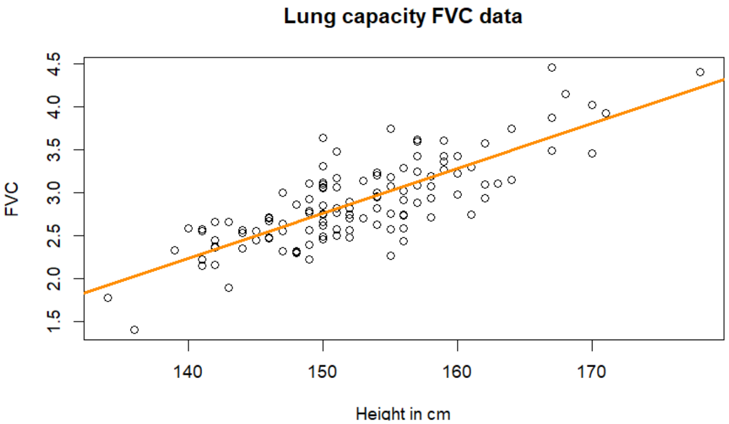
## 1. Reading the data

````
```{r}
FVC.data <- read.table("boys.dat", header = TRUE)
head(FVC.data)
```
````

| | FVC | Height | Weight |
|---|---|---|---|
| | <dbl> | <int> | <int> |
| 1 | 2.75 | 156 | 51 |
| 2 | 2.66 | 142 | 37 |
| 3 | 2.32 | 148 | 35 |
| 4 | 4.40 | 178 | 58 |
| 5 | 2.70 | 146 | 38 |
| 6 | 2.35 | 144 | 35 |

## 2. Modelling

````
```{r}
fvc.lm=lm(FVC ~ Height, data = FVC.data)
summary(fvc.lm)
plot(FVC ~ Height, data = FVC.data, xlab = "Height in
   cm",ylab = "FVC", main = "Lung capacity FVC
   data")
abline(fvc.lm, lwd = 3, col = "darkorange")
```
````

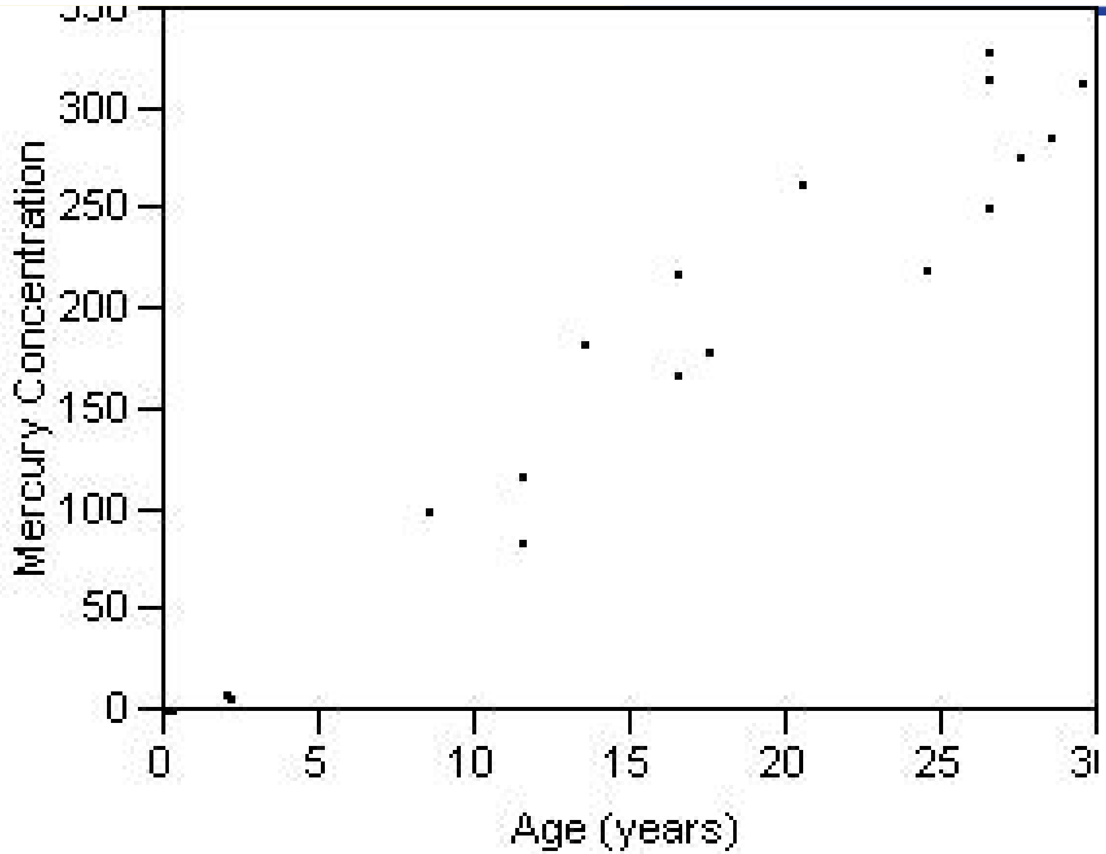



```
Call:
lm(formula = FVC ~ Height, data = FVC.data)

Residuals:
     Min       1Q   Median       3Q      Max 
-0.75507 -0.23898 -0.00411  0.21238  0.87589 

Coefficients:
             Estimate  Std. Error  t value   Pr(>|t|)
(Intercept) -5.064961   0.552593   -9.166    1.24e-15 ***
Height       0.052194   0.003618   14.426    < 2e-16 ***
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 0.3137 on 125 degrees of freedom
Multiple R-squared:  0.6248,  Adjusted R-squared:  0.6218
F-statistic: 208.1 on 1 and 125 DF,  p-value: < 2.2e-16
```

# Example 8

Dolphin example.

Data for 19 dolphins was collected as part of a marine population study. The data contains the mercury concentration (y) in the liver of striped dolphins against the age of the dolphins (x).



# Interpretation of least squares coefficients: (b0 and b1)

Mercury Concentration (µg/g) = -2.65 + 10.90 Age (years)

Slope interpretation: The mercury concentration is predicted to increase by 10.90 µg/g (slope) for every one unit year increase in age.

Intercept interpretation: Not interpretable as Age=0 is not of practical interest in this scenario.

# Aim 4 Inference about the coefficients

- The following assumptions we made about the error term are required for valid inference:
  1. $E(\epsilon_i) = 0$, for all $i$
  2. $\mathrm{var}(\epsilon_i) = \sigma^2$, for all $i$
  3. $\epsilon_i$ and $\epsilon_j$ are independent for all $i \neq j$
  4. $\epsilon_i \sim N(0, \sigma^2)$ if we wish to make inferences about the regression model
- Checking the validity of these assumptions is an important part of model-checking
- For the time being, we shall suppose that the assumptions have been satisfied

# Assuming normality

- In many situations we will be happy to assume that

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

- This means that

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

Because each parameter estimate is just a linear function of the data Y, each estimate $\hat{\beta}$ is also normally distributed

- We can place confidence intervals on the regression parameter estimates using the central limit theorem and the variance terms we calculated in the next slide.

- In order to form confidence intervals we need to know

$$Var(\hat{\beta}_0) = \sigma^2 \left( \frac{1}{n} + \frac{\overline{X}^2}{\sum (X_i - \overline{X})^2} \right); Var(\hat{\beta}_1) = \frac{\sigma^2}{\sum (X_i - \overline{X})^2}$$

- We can show this in a similar way i.e.

$$Var(\hat{\beta}_1) = Var(\sum c_i (Y_i - \overline{Y})) = Var(\sum c_i Y_i)$$

OR we can do it the easier way using matrices (later).

- These are calculated for us by R

- This is what we can calculate about least squares estimates without assuming that the errors are normally distributed. These results are used often in econometrics and other fields when it is assumed that normality of errors doesn't hold.

# Estimating the variance of the error

- In preparation for statistical inference in the linear model, we require an estimate of the error variance $\sigma^2 = \mathrm{var}(\epsilon_i)$
- Now

$$\sigma^2 = \mathrm{var}(\epsilon_i) = \mathrm{var}(y_i - \beta_0 - \beta_1 x_i)$$

  but we can only estimate the coefficients!
- It can be shown that an unbiased estimate of $\sigma^2$ is given by

$$s^2 = \frac{\mathrm{RSS}}{n-2} = \frac{1}{n-2}\sum_{i=1}^{n} \hat{e}_i^2$$

  RSS: Residual Sum of Squares
- Note that the divisor is $n - 2$ because we have estimated two parameters

# Inference about the slope

- **Sampling distribution of $\widehat{\beta}_1$.** It can be shown that if the assumptions are satisfied: $\hat{\beta}_1 \mid X \sim N(\beta_1, \frac{\sigma^2}{S_{xx}})$
- **Test statistic.** Because we have to estimate $\sigma^2$ by $s^2$, we use the statistic $T = \frac{\widehat{\beta}_1 - \beta_1^0}{s/\sqrt{S_{xx}}} = \frac{\widehat{\beta}_1 - \beta_1^0}{\text{se}(\widehat{\beta}_1)} \sim t_{n-2}$

  to test a null hypothesis $H_0: \beta_1 = \beta_1^0$; by default, *R* tests $H_0: \beta_1 = 0$ against a two-sided alternative
- **Confidence Interval.** By the same arguments, a $100(1-\alpha)\%$ confidence interval for $\beta_1$ is given by

$$\hat{\beta}_1 \pm t_{\alpha/2, n-2} \times \text{se}(\hat{\beta}_1)$$

# Significance Test for Regression Slope

## Significance Test for Regression Slope

$$\widehat{\beta_0} = b_0$$

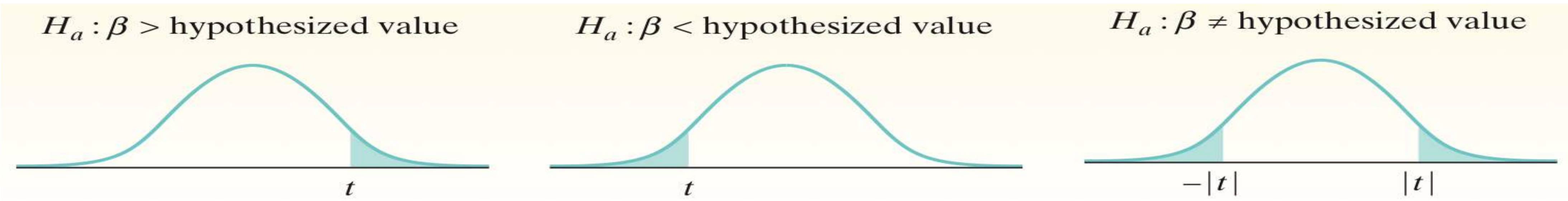
$$\widehat{\beta_1} = b_1$$

- To test the hypothesis $H_0$: $\beta_1$ = hypothesized value, compute the test statistic:

$$t = \frac{b_1 - \text{hypothesized value}}{SE_b}$$

- Find the *P*-value by calculating the probability of getting a *t* statistic this large or larger in the direction specified by the alternative hypothesis $H_a$.

- Use the *t* distribution with df = $n - 2$.



# Testing the Hypothesis of No Relationship

- We may look for evidence of a **significant relationship** between variables x and y in the population from which our data were drawn.
- For that, we can test the hypothesis that the regression slope parameter $\beta_1$ is equal to zero.

**STEP 1** $H_0$: $\beta_1 = 0$ **vs.** $H_a$: $\beta_1 \neq 0$

Testing $H_0$: $\beta_1 = 0$ is equivalent to testing the **hypothesis of no correlation** between $x$ and $y$ in the population.

**STEP 2 Test statistic** $T = \frac{\widehat{\beta}_1 - 0}{\mathrm{se}(\widehat{\beta}_1)}$

**STEP 3 The sampling distribution** $T \sim t_{n-2}$

**STEP 4 The p-value (see Ha): p-val= P(**$|t_{n-2}|$ **> t) (for two sided)**

**STEPS 5 and 6 Decision and Conclusion**

**Note:** *A test of hypothesis for* $\beta_0$ *is seldom of interest, mainly because* $\beta_0$ *often has no practical interpretation. Remember that* $\beta_0$ *represents the value of the response variable when x = 0, which is often outside the range of experimentation.*

# Inference about the intercept

- As before, it can be shown that the sampling distribution for $\hat{\beta}_0$

$$\hat{\beta}_0 \mid X \sim N(\beta_0, \sigma^2 \left( \frac{1}{n} + \frac{\bar{x}^2}{S_{xx}} \right))$$

$$\widehat{\beta_0} = b_0$$

$$\widehat{\beta_1} = b_1$$

- For hypothesis testing, we use

$$T = \frac{\hat{\beta}_0 - \beta_0^0}{s\sqrt{\frac{1}{n} + \frac{\bar{x}}{S_{xx}}}} = \frac{\hat{\beta}_0 - \beta_0^0}{\mathrm{se}(\hat{\beta}_0)} \sim t_{n-2}$$

to test the null hypothesis $H_0: \beta_0 = \beta_0^0$; by default, *R* tests $H_0: \beta_0 = 0$ against a two-sided alternative

- A $100(1 - \alpha)\%$ confidence interval for $\beta_0$ is given by

$$\hat{\beta}_0 \pm t_{\alpha/2, n-2} \times \mathrm{se}(\hat{\beta}_0)$$

# Example 9 in *R* (*function lm()*)

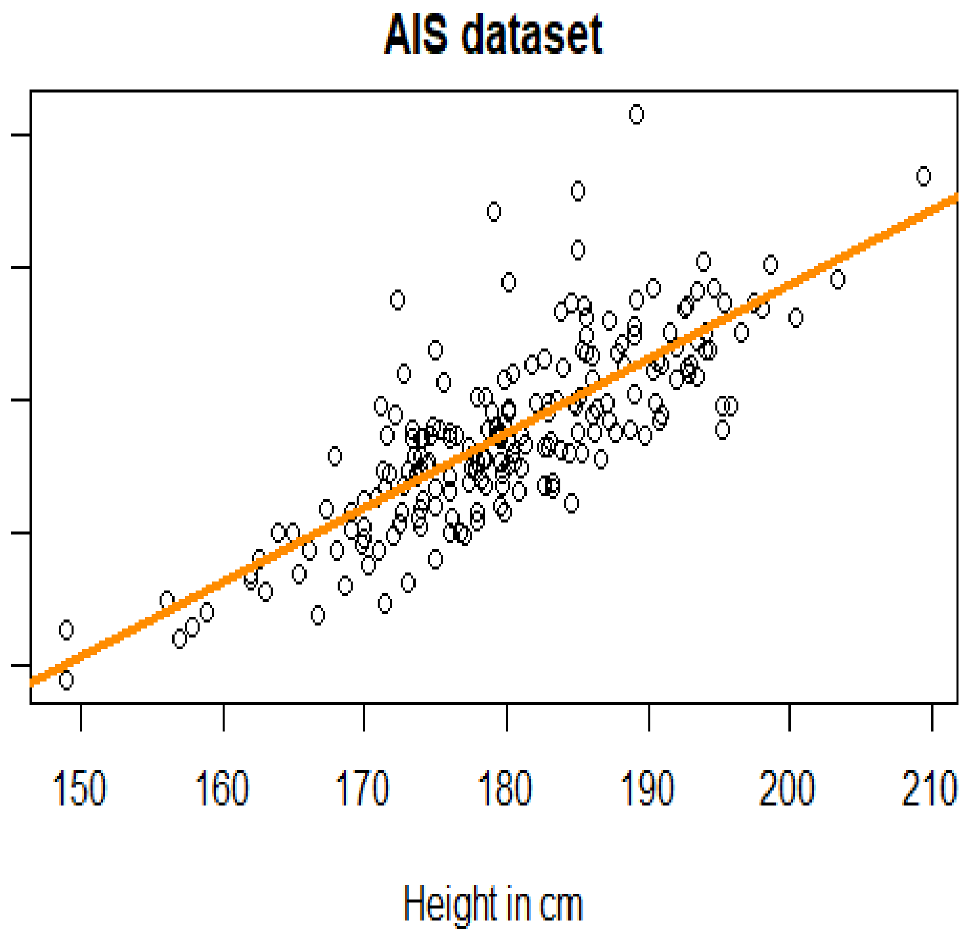
ais is a data object in *R* that contains measurements of physiological variables on 202 male and female athletes from the AIS

>ais.lm=lm(Wt ~ Ht, data = ais)

>summary(ais.lm)

>plot(Wt ~ Ht, data = ais, xlab = "Height in cm",ylab = "Weight in kg", main = "AIS dataset")

>abline(ais.lm, lwd = 3, col = "darkorange")



# Example 9: R output for athlete weight/height data

```
> summary(ais.lm)
Call:
lm(formula = Wt ~ Ht, data = ais)

Residuals:
    Min      1Q  Median      3Q     Max
-16.372  -5.296  -1.197   4.378  38.030

Coefficients:
              Estimate Std. Error t value Pr(>|t|)
(Intercept) -126.18901   11.39656  -11.07   <2e-16 ***
Ht             1.11712    0.06319   17.68   <2e-16 ***
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 8.72 on 200 degrees of freedom
Multiple R-squared:  0.6098,	Adjusted R-squared:  0.6079
F-statistic: 312.6 on 1 and 200 DF,  p-value: < 2.2e-16
```

```
> confint(ais.lm)  ## 95% CI for parameters
```

| | 2.5 % | 97.5 % |
|---|---|---|
| (Intercept) | -148.6618436 | -103.716178 |
| Ht | 0.9925209 | 1.241713 |

```
> cor(ais$Ht, ais$Wt)     0.7809063
```

# Example 9 Using R output

We want to perform a test of $H_0 : \beta_1 = 0 \quad H_a : \beta_1 > 0$

where $\beta_1$ is the true slope of the population regression line between Weight and Height on 202 male and female athletes from the AIS

STEP 1 $H_0$: $\beta_1 = 0$ vs. $H_a$: $\beta_1 > 0$

STEP 2 Test statistic $T = \frac{\widehat{\beta}_1 - 0}{\mathrm{se}(\widehat{\beta}_1)} = \frac{1.11712}{0.06319} = 17.68$

STEP 3 The sampling distribution $T \sim t_{n-2}$ that is $T \sim t_{200}$ given n=202

STEP 4 The p-value (see Ha): p-val= $P(t_{200} > 17.68)$
=pt(17.68,200,lower.tail=F)

=4.814125e-43

STEPS 5 and 6 Decision and Conclusion. As the p-value is very small, we reject the Ho. We conclude that there is a positive relationship between Weight and Height of the athletes from AIS.