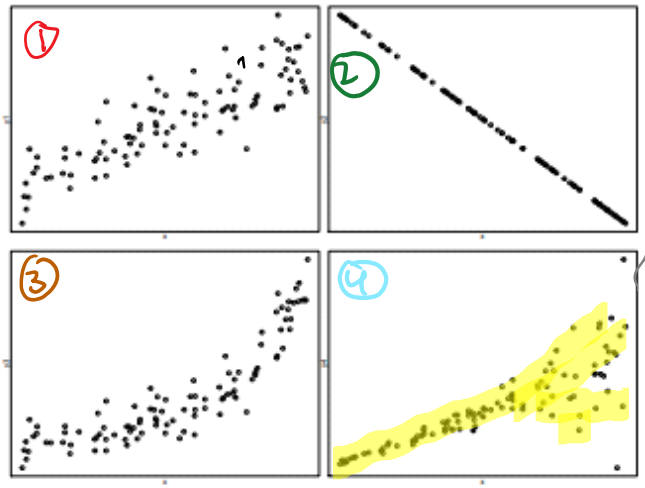
## Preliminary activity II




note: trend in difference does not matter which x you pick, has different noise, while others have the same noise.

1 Description: - increasing / positive relationship
- noise
- linear

3 Description: - increasing / positive
- has noise
- curved

2 Description: - decreasing / negative
- no noise

4 Description: - increasing / positive
- has noise

Heteroscedasticity: the circumstance in which the variability of a variable is unequal across the range of values of a second variable that predicts it.

## Regression analysis

- Statistical methodology that utilizes the relation between variables.
- Predicts a response variable (or outcome) from the relation between the response and other variables.
- Regression analysis is used in many disciplines such as:

  - Business:

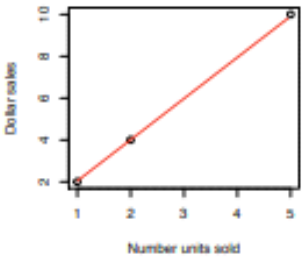
    i) Forecasting: predicting future demand for a product.

    ii) Optimization: fine tune manufacturing and delivery processes

## Functional relation

- Relation of the form
$$Y = f(X),$$
where $X$, $Y$ are variables, and $f$ is a function.
- Example: Relation between dollar sales ($Y$) of a product sales sold \$2 per unit and number of units sold ($X$):

$Y = 2x$

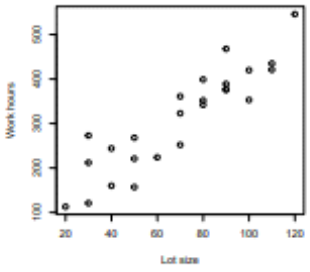



This is an example of a perfect relationship.

All observations fall on the line of functional relationship.

## Statistical relation

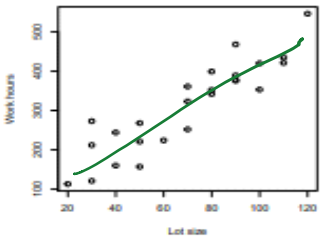
- Not a perfect relation.
- Example: A company produces replacement parts. It produces lots of varying size. The relation between the lot size and work hours is a statistical relation.




Note: There is still a trend but a greater amount of noise

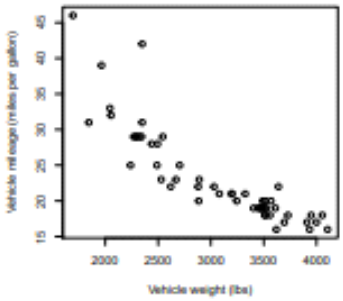
## Statistical relation

- Example (contd): There is a relation between $X$ and $Y$: the higher the lot size, the higher the work hours tend to be.
- Perfect relation?
  No! since there is noise and data points are scattered around the trend.
- Two lots with $X = 40$ have different $Y$.
- Linear or non-linear statistical relation?
  Linear statistical relation



## Statistical relation

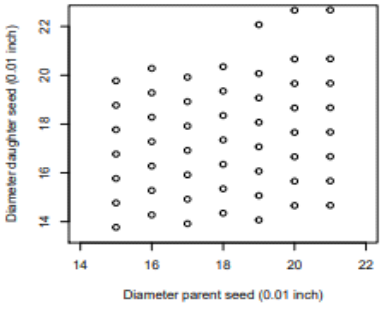
- Example: Weight and mileage for 54 cars.
- Functional or statistical relation?
- Linear statistical relation?




## Galton's early considerations of regression

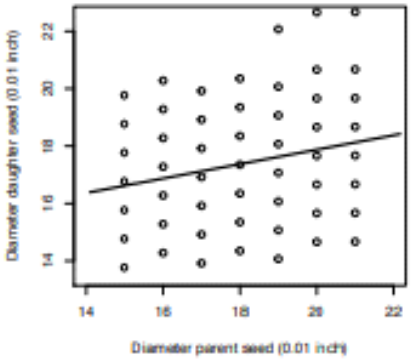
- Sir Francis Galton, English Victorian statistician, sociologist, psychologist, anthropologist, etc.
- Work on inherited characteristics of sweet peas ⇒ initial conceptualization of linear regression.
- In 1975, Galton distributed packets of sweet pea seeds to seven friends who harvested seeds from the new generations of plants and returned them to Galton.
- Galton plotted the diameter of the daughter seeds against the diameter of the mother seeds [Galton, 1894].

## Galton's early considerations of regression



## Galton's early considerations of regression

- Mean diameter of daughter seeds from a particular diameter of mother seed approximately a straight line with positive slope
  Tendency of diameter of daughter seeds to vary with diameter of mother seeds
- Constant variability for diameter of daughter seeds from a particular diameter of mother seed
  Random scatter around this tendency




## Notation and general concepts

- **Model**: mathematical expression to describe the behavior of a random variable of interest
- **Response variable** or **outcome** $Y$: variable of interest
- **Predictor** or **independent variables** $X$: know constant variables thought to provide information on the behavior of $Y$
- Subscript on $Y$ and $X$ identifies the particular unit from which the observation was taken ($X_5$ for unit 5)
- **Parameters**: control behavior of the model; usually represented by Greek letters ($\beta$, $\sigma$); unknown constants to be estimated from the sample
- **Linear model**: model linear in the parameters

Note: A model is a representation of reality. No model is a 100% accurate but can be close to matching reality.

## Examples

- Dollar sales of a product sales sold \$2 per unit and number of units sold

$$Y = \beta X$$

- Diameter of daughter seeds and diameter of mother seeds

$$Y = \beta X + \varepsilon$$

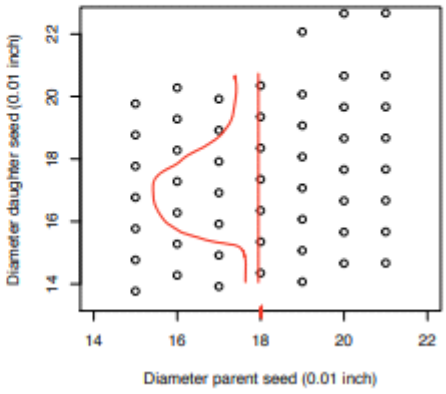
## Basic concepts

**Two characteristics of a statistical relation**:

1. Tendency of $Y$ to vary with $X$
2. Random scatter around this tendency

**In a regression model**:

1. The mean of $Y$ vary in a systematic fashion with $X$
2. Probability distribution of $Y$ for any given value of $X$



## Data collection for regression analysis

Note: observational studies can't conclude cause and effect, only correlation.

- **Observational study**
  - Investigator has no control over the explanatory variables ($X$)
  - Limitation: not adequate for cause-and-effect
    A strong association does not necessarily means a cause-and-effect relationship
- **Experiment**
  - Investigator exercises control over the explanatory variables ($X$) through random assignment
  - Random assignment balances out effect of other variables that might affect $Y$
  - Gold standard for cause-and-effect conclusions

## Example of observational study

Study the relationship between age of employees ($X$) and number of days of illness last year ($Y$)

- Observational data because we can't control age or # of sick days-
- An observed association between $X$ and $Y$ does not necessarily imply that $X$ explains $Y$

- Note: There maybe other factors that we have not looked at.

## Example of experiment

Study the relationship between productivity and length of training of analysts working in a bank:

1. 30 analysts considered
2. randomly select 10 analysts that will be trained for 2 week; randomly select 10 other analysts that will be trained for 5 weeks; the 10 remaining will be trained for 8 weeks
3. productivity of the 30 analysts observed for a fixed time after the training

- Experiment because investigators can manipulate the value of x
  - eg. 8 weeks
  - This could have cause and effect.

## Cause-and-effect / Causation

- We observe an association between $Y$ and $X$
- Does changing one of the variables imply the other to change?
- Mechanisms that can result in an observed association between $Y$ and $X$:

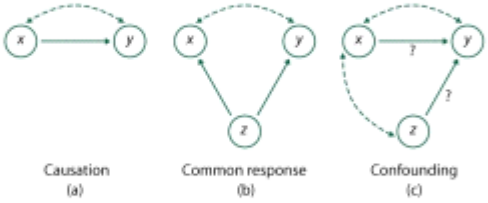



Figure 1: The dashed arrows represent association and the solid ones cause and effect link. The variable $x$ is explanatory, $y$ is response, and $z$ is a lurking variable.

Regression analysis by itself provides no information about causation. Be careful in drawing causal conclusions

## Overview of the steps in regression analysis

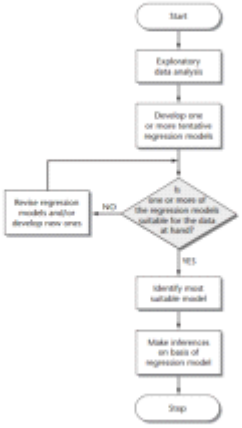



Figure 2: The steps in regression analysis [Kutner et al., 2004, p.14]

## Three main purposes of regression analysis

1. **Describe**: describe the relation between diameter of daughter seeds and diameter of mother seeds.
2. **Control**: control the length of training to maximize productivity constrained by costs.
3. **Predict**: predict future demand for a product.

## Parameters, estimators, and estimates

- **Parameter**: quantity of interest, quantity describing a population (or model).
  A parameter is a constant (constant/random) quantity.
- **Estimator**: rule for calculating an estimate of parameter.
  An estimator is a random (constant/random) quantity.
- **Estimate**: result of the estimator (for a given sample).
  An estimate is a constant (constant/random) quantity.


Recall

$\bar{x} = \frac{1}{N} \sum x_i$

## Toluca company example[1]

- Toluca Company produces replacement parts for refrigeration equipment
- Produces lots of varying size
- Cost improvement: find optimal lot size
- Key input: relationship between lot size and labor hours
- Data: lot size $X$ and work hours $Y$ for 25 production runs

| Run $i$ | Lot size $X_i$ | Work hours $Y_i$ |
|---|---|---|
| 1 | 80 | 399 |
| 2 | 30 | 121 |
| ... | ... | ... |
| 24 | 80 | 342 |
| 25 | 70 | 323 |

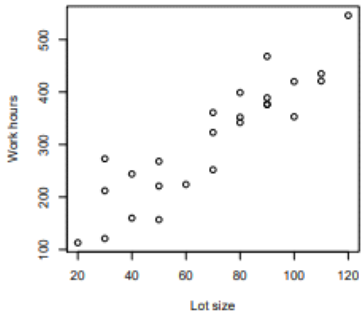
← how much time it takes to produce the items

ex / 399 hours to create 80 fridges

[1] From [Kutner et al., 2004], page 19

## Toluca company example

From the scatter plot:
- looks like a linear model




## Simple linear model

Suppose we have $n$ observed pairs $(X_i, Y_i)$, $i = 1, \ldots, n$. The simple linear model is:

$$Y_i = \beta_0 + \beta_1 X_i + \varepsilon_i,$$

where

4 assumptions

- $Y_i$ is the observed value of $Y$ on unit $i$,
- $\beta_0$ and $\beta_1$ are parameters,
- $X_i$ is the observed value of $X$ on unit $i$, and
- $\varepsilon_i$ are random errors that have zero mean $\mathrm{E}(\varepsilon_i) = 0$, with common variance $\mathrm{Var}(\varepsilon_i) = \sigma^2$, and pairwise independent.

$\varepsilon_i \perp \varepsilon_j, \; i \neq j$

## Simple linear model

### Exercise 1

*Show that the random errors satisfy*

$$E(\varepsilon_i\varepsilon_j) = \begin{cases} 0 & \text{if } i \neq j \\ \sigma^2 & \text{if } i = j \end{cases}$$

Recall Assumptions about random errors

1) $E(\varepsilon_i) = 0$
2) $Var(\varepsilon_i) = \sigma^2$
3) pairwise independent; thus $Cov(\varepsilon_i, \varepsilon_j) = 0 \quad i \neq j$

For this proof there are 2 cases $i=j$ & $i \neq j$

- $i=j$:

wts $E(\varepsilon_i \cdot \varepsilon_i) = E(\varepsilon_i^2) = \sigma^2$

We know: $Var(\varepsilon_i) = \sigma^2$

So $Var(\varepsilon_i) = E(\varepsilon_i^2) - (E(\varepsilon_i))^2$ , first assumption says $E(\varepsilon_i)=0$ , second assumption says $Var(\varepsilon_i)=\sigma^2$

$\sigma^2 = \sigma^2 - 0$

$\sigma^2 = \sigma^2$

Recall
$Var(x) = E(x^2) - (E(x))^2$

- $i \neq j$: we want to show $E(\varepsilon_i \varepsilon_j) = 0$

$0 = E(\varepsilon_i \varepsilon_j) - E(\varepsilon_i)E(\varepsilon_j)$ , first assumption

$0 = E(\varepsilon_i \varepsilon_j)$

Recall
$Cov(\varepsilon_i, \varepsilon_j) = 0$

## Important features

Simple linear model

constant $\Rightarrow$ understand constant as not random.

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

Parameters are always constant. We don't know them but they are constant

1. The response $Y_i$ is a sum of two terms:
   - A constant term
   - A random term

   The outcome $Y_i$ is random (constant/random)
2. $E(Y_i) = \beta_0 + \beta_1 X_i$, where $E(Y_i)$ is a shortcut for $E(Y_i|X_i)$ the mean of $Y$ when $X = X_i$.

   $Y_i$ is constant + random = random

   $E(y_i) = E(\beta_0 + \beta_1 x_i + \varepsilon_i) = E(\beta_0) + E(\beta_1 x_i) + E(\varepsilon_i)$ , linearity

   $= \beta_0 + \beta_1 x_i$

   Thus, the functional relationship between the true mean of $Y_i$ and $X_i$ is a straight line with intercept $\beta_0$ and slope $\beta_1$

## Important features

Simple linear model

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

3. $\mathrm{Var}(Y_i) = \sigma^2$, where $\mathrm{Var}(Y_i)$ is a shortcut for $\mathrm{Var}(Y_i|X_i)$ the ~~mean~~ variance of $Y$ when $X = X_i$.

$\mathrm{Var}(y_i) = \mathrm{Var}(\underbrace{\beta_0 + \beta_1 X_i}_{\text{constant}} + \varepsilon_i) = \mathrm{Var}(\varepsilon_i) = \sigma^2$

4. The outcomes $Y_i$ are pairwise independent because the errors $\varepsilon_i$ are pairwise independent.

i.e $y_i \perp$ from $y_j$ when $i \neq j$

Recall
- Variance is not linear.
- You could use expectation but too hard.
- $\beta_0 + \beta_1 x_i$ is a constant

Var (constant + r.v) = var(r.v)

## Reminder: normal distribution

A random variable $X$ is normal if its probability density function is

$$f_X(x) = \frac{1}{\sigma\sqrt{2\pi}} \exp\left\{-\frac{1}{2}\left(\frac{x-\mu}{\sigma}\right)^2\right\},$$

where $\mu \in \mathbb{R}$ and $\sigma \in \mathbb{R}^+$ are the parameters of the distribution. We say that $X$ is normally distributed with mean $\mathrm{E}(X) = \mu$ and variance $\mathrm{Var}(X) = \sigma^2$ and we write

$$X \sim \mathcal{N}(\mu, \sigma^2).$$

Additional Assumption

Sometimes we make an additional assumption that random errors are normally distributed.

- This assumption is only used when explicitly stated

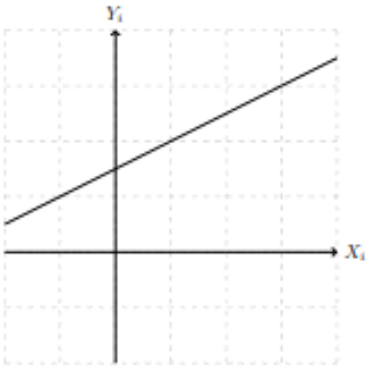
## Simple linear model with normal errors

- The random errors are sometimes assumed to be normally distributed.
- Simple linear model with normal errors:

$$Y_i = \beta_0 + \beta_1 X_i + \varepsilon_i,$$

where
  - $\beta_0$ and $\beta_1$ are parameters,
  - $\varepsilon_i$ are independently and identically distributed (i.i.d.) with normal distribution with mean 0 and variance $\sigma^2$.
- In what follows, we suppose a simple linear model (errors not necessarily normal) unless otherwise specified.

## Interpretation of the regression parameters




- If the scope of the model includes $X = 0$, the intercept $\beta_0$ is the mean of $Y$ when $X = 0$ (no meaning otherwise)
- The slope $\beta_1$ is the change in the mean of $Y$ per unit increase of $X$

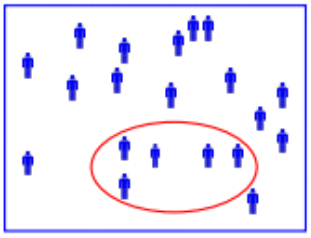
## Estimation of the parameters

- Postulated model:

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

- Observed values $(X_i, Y_i)$
- Parameters $\beta_0$ and $\beta_1$ unknown and to be estimated from the sample.
- Two estimation methods:
  1. Least squares
  2. Maximum likelihood

  $\Rightarrow$ Estimates $\hat{\beta}_0$ and $\hat{\beta}_1$



- See the model as describing the population
- Select sample at random
- observe x & y values
- estimate $\beta_0$ and $\beta_1$ from the sample

^ <- that means estimate, not true values.

## Method of least squares

Simple linear model:

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

- Parameters $\beta_0$ and $\beta_1$ to be estimated from the data.
- **Goal**: find the best estimates $\hat{\beta}_0$ and $\hat{\beta}_1$ given the data.
- What does *best* mean?
- **Least square**: best by criterion

$$Q(\beta_0, \beta_1) = \sum_{i=1}^{n} (Y_i - \beta_0 - \beta_1 X_i)^2$$

$y_i - \beta_0 - \beta_1 y_i$ is the deviation of $Y_i$ from its expected value. $y_i - E(y_i)$

- Least square estimators of $\beta_0$ and $\beta_1$: $\hat{\beta}_0$ and $\hat{\beta}_1$ that minimize criterion $Q$.

We want to minimize the sum of the errors to get the best fit

## Least square estimators

Find $\hat{\beta}_0$ and $\hat{\beta}_1$ that minimize criterion

$$Q(\beta_0, \beta_1) = \sum_{i=1}^{n} (Y_i - \beta_0 - \beta_1 X_i)^2$$

given the data.

1. Write the normal equations (derivatives of $Q$ set to 0).
2. Find the critical points (solution of the normal equations).
3. Determine whether the critical point is a maximum or a minimum (we will skip this step).

1) $\frac{\partial Q}{\partial \beta_0} = 2 \sum_{i=1}^{n} (y_i - \beta_0 - \beta_1 x_i) \quad (-1)$

$0 = -2 \sum_{i=1}^{n} (y_i - \beta_0 - \beta_1 x_i)$

$\frac{\partial Q}{\partial \beta_1} = 2 \sum_{i=1}^{n} (y_i - \beta_0 - \beta_1 x_i) \cdot (-x_i)$

$0 = 2 \sum_{i=1}^{n} (y_i - \beta_0 - \beta_1 x_i) \cdot (-x_i)$

2) $\beta_0$ : Critical points

$0 = -2 \sum_{i=1}^{n} (y_i - \beta_0 - \beta_1 x_i)$

$0 = \sum_{i=1}^{n} y_i - \sum_{i=1}^{n} \beta_0 - \beta_1 \sum_{i=1}^{n} x_i$

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

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

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

$\hat{\beta}_0 = \bar{y} - \beta_1 \bar{x}$

$\beta_1$: Critical point

$0 = 2 \sum_{i=1}^{n} (y_i - \beta_0 - \beta_1 x_i) \cdot (-x_i)$

$0 = \sum_{i=1}^{n} (y_i x_i - \beta_0 x_i - \beta_1 x_i^2)$

$0 = \sum_{i=1}^{n} y_i x_i - \beta_0 \sum_{i=1}^{n} (x_i) - \beta_1 \sum_{i=1}^{n} x_i^2$

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

$\hat{\beta}_1 = \frac{\sum_{i=1}^{n} y_i x_i - \beta_0 \sum_{i=1}^{n} x_i}{\sum_{i=1}^{n} x_i^2}$

## Least square estimators

Least square estimators of $\beta_1$ and $\beta_0$:

$$\hat{\beta}_1 = \frac{\sum_{i=1}^{n} X_i Y_i - \frac{1}{n} \sum_{i=1}^{n} X_i \sum_{i=1}^{n} Y_i}{\sum_{i=1}^{n} X_i^2 - \frac{1}{n} \left( \sum_{i=1}^{n} X_i \right)^2},$$

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

where

- $\overline{Y} = \frac{1}{n} \sum_{i=1}^{n} Y_i$ and
- $\overline{X} = \frac{1}{n} \sum_{i=1}^{n} X_i$

are the sample mean of $Y$ and $X$, respectively.

## Least square estimators

### Exercise 2

*Show that*

$$\hat{\beta}_1 = \frac{\sum_{i=1}^{n}(X_i - \overline{X})(Y_i - \overline{Y})}{\sum_{i=1}^{n}(X_i - \overline{X})^2}.$$

## Regression equation

Regression equation or fitted regression line

$$\hat{Y} = \hat{\beta}_0 + \hat{\beta}_1 X$$

where $\hat{Y}$ is the estimated mean of the response variable at level $X$ of the explanatory.

## Gauss-Markov theorem

**Theorem 1**
*Consider the simple linear model*

$$Y_i = \beta_0 + \beta_1 X_i + \varepsilon_i.$$

*Suppose that the following assumptions concerning the random errors (called Gauss-Markov assumptions) are satisfied:*

- *They have mean zero:* $E(\varepsilon_i) = 0$,
- *They are homoscedastic:* $\mathrm{Var}(\varepsilon_i) = \sigma^2 < \infty$, *and*
- *There are uncorrelated* $Cov(\varepsilon_i, \varepsilon_j) = 0, \forall i \neq j$.

*Then the least squares estimators* $\widehat{\beta}_0$ *and* $\widehat{\beta}_1$ *are unbiased and have minimum variance among all unbiased linear estimators.*

## Proof of the Gauss-Markov theorem

Step 1 Exercise: Prove that the least squares estimators are unbiased, i.e. prove that

$$\mathrm{E}(\widehat{\beta}_1) = \beta_1 \quad \text{and} \quad \mathrm{E}(\widehat{\beta}_0) = \beta_0$$

Step 2 To be proven later: The least squares estimators have minimum variance among all unbiased linear estimators.

## Toluca company example

Using R, we find:

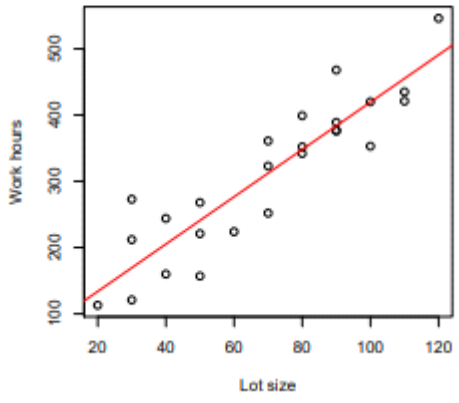
$$\sum_{i=1}^{n} X_i = 1750 \quad \sum_{i=1}^{n} Y_i = 7807 \quad \sum_{i=1}^{n} X_i Y_i = 617180$$

$$\sum_{i=1}^{n} X_i^2 = 142300 \quad n = 25$$

### Exercise 3

1. *Compute the least squares estimates of $\beta_1$ and $\beta_0$.*
2. *What is the regression equation?*
3. *Interpret the parameters.*

## Toluca company example



## Toluca company example: R output

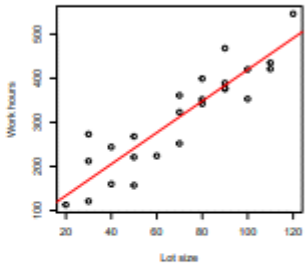
```
Call:
lm(formula = Hours ~ Size, data = toluca)

Residuals:
    Min      1Q  Median      3Q     Max
-83.876 -34.088  -5.982  38.826 103.528

Coefficients:
            Estimate Std. Error t value Pr(>|t|)
(Intercept)   62.366     26.177   2.382   0.0259 *
Size           3.570      0.347  10.290 4.45e-10 ***
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 48.82 on 23 degrees of freedom
Multiple R-squared:  0.8215,Adjusted R-squared:  0.8138
F-statistic: 105.9 on 1 and 23 DF,  p-value: 4.449e-10
```

## Preliminary exercise: Toluca company example

- Regression equation (or fitted regression line)

$$\hat{Y} = 62.37 + 3.5702X,$$

where
  - $X$ is the lot size, and
  - $Y$ is the work hours.




What is:

- The predicted work hours for a new production run for a lot size of 60?
- The estimated population mean work hours for a lot size of 60?

## Predicted values and residuals

- Fitted regression line:

$$\widehat{Y} = \widehat{\beta}_0 + \widehat{\beta}_1 X$$

- Fitted value: value of $Y$ computed from the regression line:

$$\widehat{Y}_i = \widehat{\beta}_0 + \widehat{\beta}_1 X_i.$$

  Fitted value $\widehat{Y}_i$ used as:
  - **Prediction** of the value of $Y$ for particular value $X_i$ of $X$. Sometimes written $\widehat{Y}_{pred_i}$.
  - **Estimate** of the population mean of $Y$ for particular value $X_i$ of $X$.
- A residual is the deviation of the observed value $Y_i$ from the fitted value $\widehat{Y}_i$

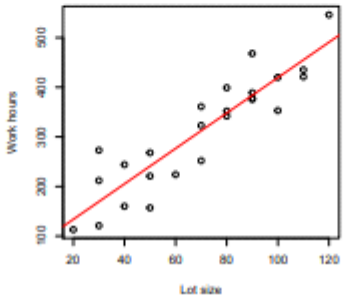
$$e_i = Y_i - \widehat{Y}_i$$

## Toluca company example

For production run 6, the lot size was 60 and 224 work hours were required.

### Exercise 4

1. *What is the fitted value for this observation?*
2. *What is the residual?*
3. *Where can we read the observed work hours ($Y$), the fitted work hours, and the residual in the scatterplot?*



## Exercises

- From the textbook[2]
  - 1.20
  - 1.21
- From the slides[3]
  - Slide 26
  - Slide 37
  - Slide 40
  - Slide 41
  - Slide 47

[2] Solutions in the student manual (CD) provided with the textbook
[3] Partial solutions posted on Quercus