AMA3602

Applied Linear Models for Finance Analytics

Department of Applied Mathematics

Lecturer : Dr. Catherine Liu

Office & Consultation venue: TU830

Consultation: 15: 45- 16: 45, Monday

Lecture time: 11:30-13:20, Tuesday, M301

16, 23, 30/01/2024

Chapter **1**

Simple Linear Regression

INTRODUCTION & SIMPLE LINEAR REGRESSION

Regression and Model Building

- **Regression analysis**:
  - a **statistical technique** for analyzing **multifactor** data, and investigating and **modeling the relationship between variables**.
  - Its broad appeal and usefulness **result from** the conceptually logical process of using an equation to express the relationship between a variable of interest (the response) and a set of related predictor variables.
  - Numerous **applications** in almost every field:
    - Engineering
    - Physical and chemical sciences
    - Economics
    - Management
    - Life and biological sciences
    - Social sciences

Data to be Analyzed

- **Data**
  - **Data collection**: n observations

$$(y_1, x_1),\ (y_2, x_2),\ \cdots\ ,\ (y_n, x_n);$$

$$\text{or} \quad (y_i, x_i), \qquad i = 1, 2, 3, \ldots, n$$

  - **Data framework**:

$$(\mathbf{y}, \mathbf{x})_{n\times 2} \equiv \begin{pmatrix} y_1\ x_1 \\ y_2\ x_2 \\ \vdots \\ y_n\ x_n \end{pmatrix}, \text{ where } \mathbf{y}_{n\times 1} = \begin{pmatrix} y_1 \\ y_2 \\ \vdots \\ y_n \end{pmatrix}, \ \mathbf{x}_{n\times 1} = \begin{pmatrix} x_1 \\ x_2 \\ \vdots \\ x_n \end{pmatrix}$$

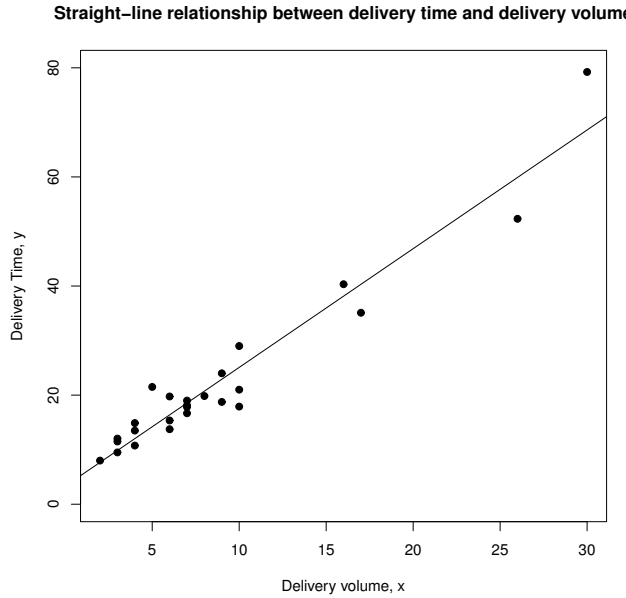
- **Aim**
  - to model the relationship between y and x;
  - to examine how x affects y.

Example 1.1: Soft Drink Beverage Data

Table: Example 1.1

Observation	Delivery_Time_y	Number_of_Cases_x
1	16.68	7
2	11.50	3
3	12.03	3
4	14.88	4
5	13.75	6
6	18.11	7
7	8.00	2
8	17.83	7
9	79.24	30
10	21.50	5
11	40.33	16
12	21.00	10
13	13.50	4
14	19.75	6
15	24.00	9
16	29.00	10
17	15.35	6
18	19.00	7
19	9.50	3
20	35.10	17
21	17.90	10
22	52.32	26
23	18.75	9
24	19.83	8
25	10.75	4

- **Background**: The engineer of a soft drink beverage bottler visits 25 randomly chosen retail outlets having vending machines, and the in-outlet delivery time (in minutes) and the volume of product delivered (in cases) are observed for each and shown in table example 1.1.
- y: Delivery time
  x: Number of cases (delivery volume)
- **Question arising**: Is the delivery time (y) related to the number of cases of product delivered (x)?
- **Scatter diagram** (scatterplot or scatter plot): a **visualization** of the relationship between two variables measured on the same set of individuals.




```
#Input data from csv file
data <- read.csv("example1.1.csv")

#Figure 1.1a (left), plot the scatter diagram
plot(data$Number.of.Cases_x,data$Delivery_Time_y, pch=19,ylim
     = c(0,80),xlab="Delivery volume, x", ylab = "Delivery Time
     , y", main="Scatter diagram for delivery volume")

#Figure 1.1b (right), pch=19 for solid points
plot(data$Number.of.Cases_x,data$Delivery_Time_y, pch=19,ylim
     = c(0,80),xlab="Delivery volume, x", ylab = "Delivery Time
     , y", main="Straight-line relationship between delivery
     time and delivery volume")

#Add straight-line relationship
abline(lm(data$Delivery_Time_y~data$Number.of.Cases_x))
```

Parameters and Variables

- **Observed data**

  $(y_i, x_i), \quad i = 1, 2, 3, \ldots, n$

- **Unknown parameters**

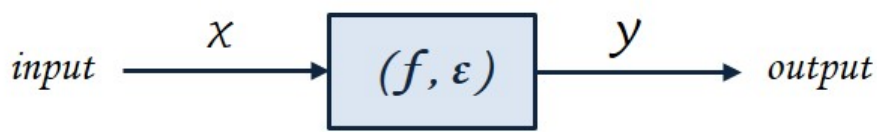
  $\beta_0 \quad \beta_0 \quad \sigma^2$

- **Random variables**

  $\varepsilon \quad y$

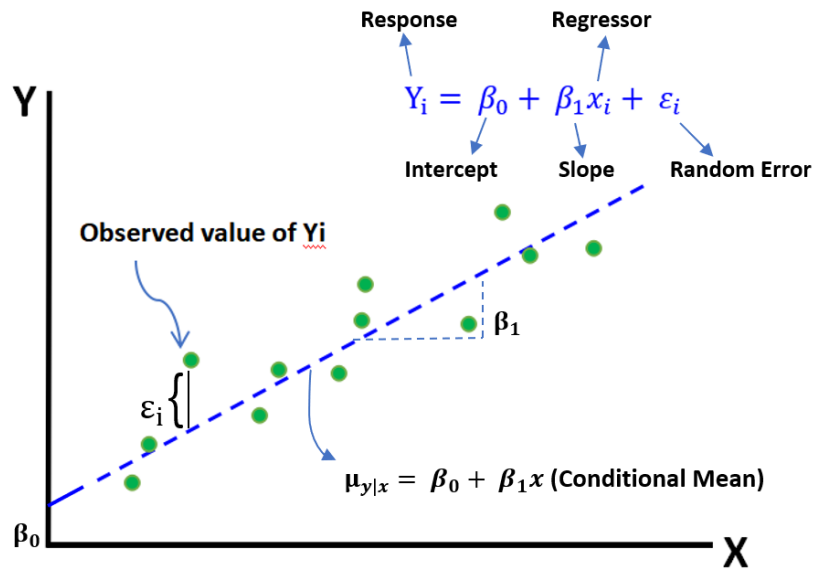
- Remark:
  - "Random" $\neq$ "Unkown"
  - In most situations, σ^2 is unknown.
  - Given the explanatory variable x, the randomness of y comes from the randomness of ε.

IMPORTANT!!



SLR Model

- The simplest regression model is **simple linear regression model** (SLR).
- **SLR model**: $y = \beta_0 + \beta_1 x + \varepsilon$, where
  - y: response variable (**dependent**)
  - β_0: intercept; β_1: slope
  - x: predictor or regressor variable (**independent**)
  - ε: **random error** with mean 0 and variance σ^2 (**no specified distribution**)



- **Model + Data** $\Rightarrow y_i = \beta_0 + \beta_1 x_i + \varepsilon, \quad i = 1, 2, 3, \ldots, n$

Mechanistic Models vs. Empirical Models

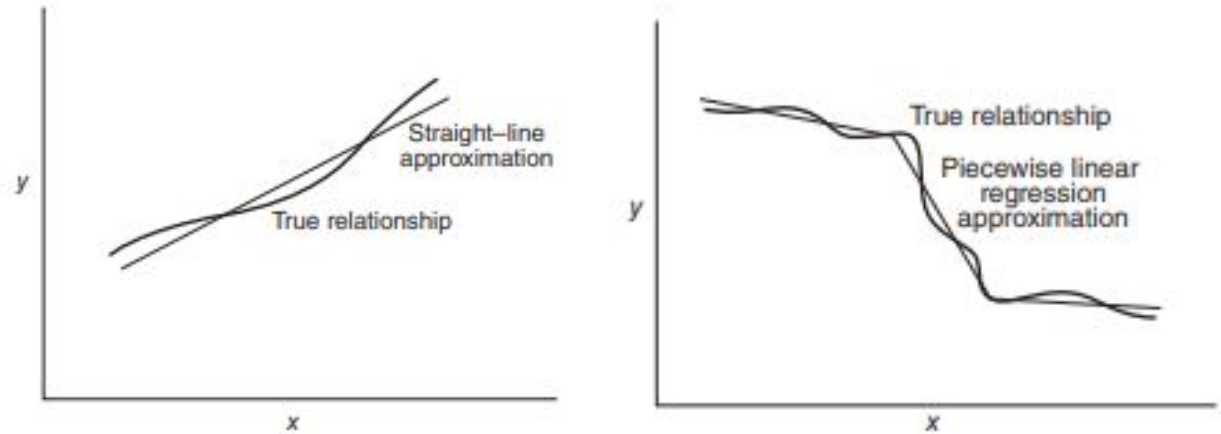
- **Mechanistic models (true relationship $\Rightarrow$ underlying mechanism)**

$$\begin{cases} y = \beta_0 + \beta_1 x + \varepsilon & \text{(with intercept)} \\ y = \beta_1 x + \varepsilon & \text{(through origin)} \end{cases}$$

- **Empirical models (approximation based on data observation)**

$$\begin{cases} y_i = \beta_0 + \beta_1 x_i + \varepsilon & \text{(with intercept)} \\ y_i = \beta_1 x_i + \varepsilon & \text{(through origin)} \end{cases} \qquad i = 1, \ldots, n$$

  - Model with intercept $\Leftarrow$ Model through origin with **location swift**



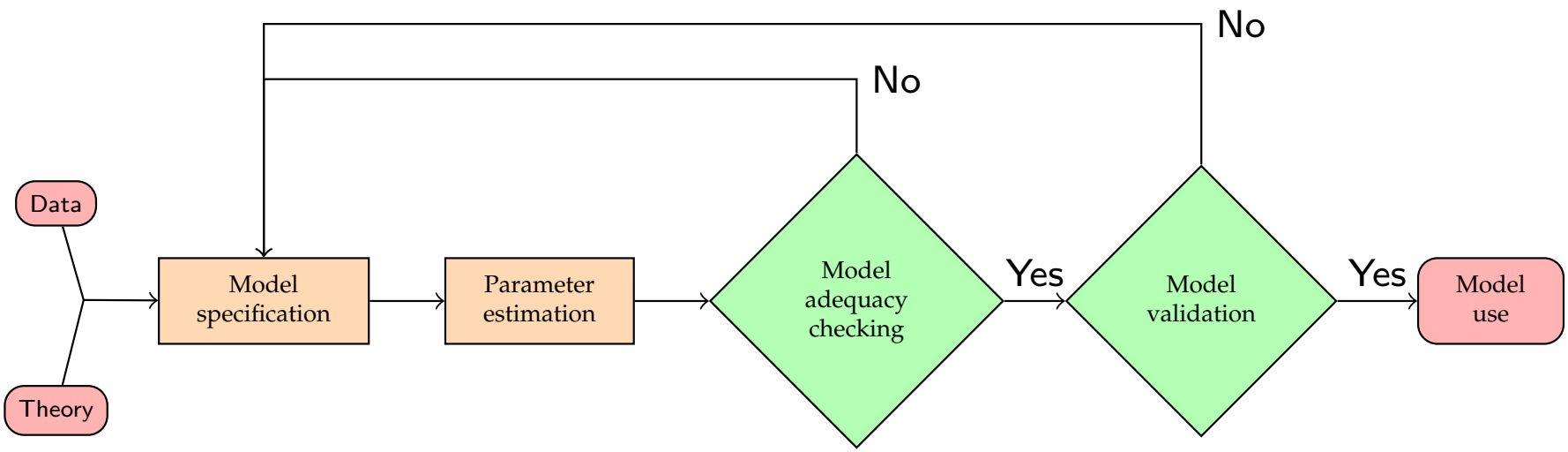
- **Remark**:
  - An important objective of regression analysis is to **estimate the unknown parameters** in the regression model. This process is also called **fitting the model to the data**.
  - The next phase of a regression analysis is called **model adequacy checking**, in which the appropriateness of the model is studied and the quality of the fit ascertained.
  - A regression model **does not** imply a **cause-and-effect relationship** between the variables.
  - The regression equation itself may not be the primary objective of the study, but it is more important to gain insights and understanding concerning the system generating the data.

Data Collection and Uses of Regression

- **Data collection**
  - Retrospective study;
  - Observational study;
  - Designed experiment.
- **Uses of regression**
  - Data description;
  - Parameter estimation;
  - Prediction and estimation;
  - Control.

Role of the Computer

Computer-aged statistics



Install **R-language**: `https://cran.r-project.org/mirrors.html`

RStudio: `https://rstudio.com/products/rstudio/download/`

Rmarkdown (can include output naturally)

SLR Model for Data $(y_i, x_i)_{i=1}^n$

- **SLR** model: $y = \beta_0 + \beta_1 x + \varepsilon$
- **Interpretation of parameters and variables**:
  - β_0 (intercept) and β_1 (slope): **regression coefficients**
  - β_0: the mean of the distribution of the response y when x = 0 if the range of the data on x includes origin, otherwise no practical interpretation.
  - β_1: the change in the mean of the distribution of y produced by a unit change in x.
  - ε: **model error**
  - y: response or dependent variable
  - x: predictor, regressor, covariate or independent variable
- Remark: The regressor x can be random. Hence we assume $x \perp \varepsilon$. The regressor variable x can be fixed. That is,

  $\cdot|X = x$: given that the regressor variable X equals to a fixed value x.
- **Model + Data** $\Rightarrow y_i = \beta_0 + \beta_1 x_i + \varepsilon, \quad i = 1, 2, 3, \ldots, n$

- **Assumptions** for SLR:

Assumption 1 : $\beta_0 + \beta_1 x = (1,\ x) \begin{pmatrix} \beta_0 \\ \beta_1 \end{pmatrix} \overset{x_0 \equiv 1}{=} \sum_{i=0}^{1} \beta_i x_i$

There is a linear relationship between y and x (linearity).

Assumption 2 : $E(\varepsilon) = 0$ (The system has **no bias**.)

- **Probabilistic view**:
  - **Mean function (Conditional mean regression)**

    $E(Y|x) = E(Y|X = x) = E(\beta_0 + \beta_1 x + \varepsilon|X = x) = \beta_0 + \beta_1 x + E(\varepsilon|X = x) \overset{x \perp \varepsilon}{=} \beta_0 + \beta_1 x + E(\varepsilon) = \beta_0 + \beta_1 x$
  - **Variance function**

    $\mathrm{Var}(Y|x) = \mathrm{Var}(Y|X = x) = \mathrm{Var}(\beta_0 + \beta_1 x + \varepsilon|X = x) = 0 + \mathrm{Var}(\varepsilon|X = x) \overset{x \perp \varepsilon}{=} 0 + \mathrm{Var}(\varepsilon) := \sigma^2$

    (The variance function is a constant.)

- Remark: $:=$ or $\equiv$ can be read "defined as".

Targets & Assumptions

- **Data observed**: $(y_i, x_i), \quad i = 1, \ldots, n$
- **Target 1**: to estimate β_0 and β_1

  Assumptions needed: Assumptions 1 & 2 plus

Assumption 3 : $\mathrm{Cov}(\varepsilon_i, \varepsilon_j) = \begin{cases} \sigma^2, & i = j \\ 0, & i \neq j \end{cases}, \quad i = 1, \ldots, n$

- **Target 2**: to estimate σ^2 (since it is unknown in most situations)

  Assumptions needed: Assumptions 1, 2 & 3

Assumption 4 : $\varepsilon_i \sim N(\cdot, \cdot), \quad i = 1, \ldots, n$

- **Target 3**: hypothesis testing on β_0 and β_1

  Assumptions needed: Assumptions 1, 2, 3 & 4
- **Target 4**: confidence interval of β_0, β_1 and mean response

  Assumptions needed: Assumptions 1, 2, 3 & 4
- **Target 5**: prediction interval of new observations

  Assumptions needed: Assumptions 1, 2, 3 & 4
- **Remark**: For any pairs of random variables X and Y,

  independent $\Rightarrow$ correlated and correlated $\not\Rightarrow$ independent.

  However, correlated $\overset{X,Y \sim N(\cdot,\cdot)}{\Longleftrightarrow}$ independent.

Target 1: Estimation of β_0 and β_1

- n pairs of sample data: $(y_i, x_i), \quad i = 1, \ldots, n$
- **Sample SLR model**: $y_i = \beta_0 + \beta_1 x_i + \varepsilon_i, \quad i = 1, \ldots, n$
- **Assumptions**: Assumptions 1 & 2
- **Least-squares estimation (LSE) method**: we estimate β_0 and β_1 so that the sum of squares of the differences between the observations y_i and the straight line $\beta_0 + \beta_1 x_i$ is a minimum.
- $(\hat{\beta}_0, \ \hat{\beta}_1)$ is the least-square estimators of $(\beta_0, \ \beta_1)$.
- The least-squares estimators $(\hat{\beta}_0, \ \hat{\beta}_1)$ **minimize**

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

- $\hat{\beta}_0$ and $\hat{\beta}_1$ satisfy

$$\left.\frac{\partial S}{\partial \beta_0}\right|_{\hat{\beta}_0,\ \hat{\beta}_1} = -2\sum_{i=1}^{n}(y_i - \hat{\beta}_0 - \hat{\beta}_1 x_i) = 0$$

$$\left.\frac{\partial S}{\partial \beta_1}\right|_{\hat{\beta}_0,\ \hat{\beta}_1} = -2\sum_{i=1}^{n}(y_i - \hat{\beta}_0 - \hat{\beta}_1 x_i)x_i = 0$$

- Simplifying these equations yields the **least-squares normal equations**:

$$n\hat{\beta}_0 + \hat{\beta}_1 \sum_{i=1}^{n} x_i = \sum_{i=1}^{n} y_i$$

$$\hat{\beta}_0 \sum_{i=1}^{n} x_i + \hat{\beta}_1 \sum_{i=1}^{n} x_i^2 = \sum_{i=1}^{n} y_i x_i$$

- The solution to the normal equations is

$$\hat{\beta}_0 = \bar{y} - \hat{\beta}_1 \bar{x} \qquad \hat{\beta}_1 = \frac{S_{xy}}{S_{xx}} = \left(\frac{S_{xy}/n}{S_{xx}/n}\right)$$

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

$S_{xx} = \sum_{i=1}^{n}(x_i - \bar{x})^2$ and $S_{xy} = \sum_{i=1}^{n} y_i(x_i - \bar{x})$.

- S_{xx} is the **corrected sum of squares** of the x_i.

  $S_{xx} = \sum_{i=1}^{n}(x_i - \bar{x})^2 = \sum_{i=1}^{n}(x_i^2 - 2x_i\bar{x} + \bar{x}^2) = \sum_{i=1}^{n} x_i^2 - 2\bar{x}\sum_{i=1}^{n} x_i + \sum_{i=1}^{n}\bar{x}^2$

  $= \sum_{i=1}^{n} x_i^2 - 2n\bar{x}^2 + n\bar{x}^2 = \sum_{i=1}^{n} x_i^2 - n\bar{x}^2$

  - Corrected sum of squares: the sum of squares of the deviations of the values about their mean.

- S_{xy} is the **corrected sum of cross products** of x_i and y_i.

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

  $= \sum_{i=1}^{n} x_i y_i - \bar{y}\sum_{i=1}^{n} x_i - \bar{x}\sum_{i=1}^{n} y_i + \sum_{i=1}^{n}\bar{x}\bar{y} = \sum_{i=1}^{n} x_i y_i - n\bar{x}\bar{y} - n\bar{x}\bar{y} + n\bar{x}\bar{y}$

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

- The **fitted simple linear regression model** is then

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

- **Residual**: the difference between the observed value y_i and the corresponding fitted value $\hat{y}_i$

$$e_i = y_i - \hat{y}_i = y_i - (\hat{\beta}_0 + \hat{\beta}_1 x_i), \quad i = 1, \ldots, n$$

Residuals play an important role in investigating **model adequacy** and in detecting **departures** from the underlying assumptions.

Example 1.2: The Rocket Propellant Data

Table: Example 1.2

Observation	Shear Strength, y	Age of Propellant, x
1	2158.70	15.50
2	1678.15	23.75
3	2316.00	8.00
4	2061.30	17.00
5	2207.50	5.50
6	1708.30	19.00
7	1784.70	24.00
8	2575.00	2.50
9	2357.90	7.50
10	2256.70	11.00
11	2165.20	13.00
12	2399.55	3.75
13	1779.80	25.00
14	2336.75	9.75
15	1765.30	22.00
16	2053.50	18.00
17	2414.40	6.00
18	2200.50	12.50
19	2654.20	2.00
20	1753.70	21.50

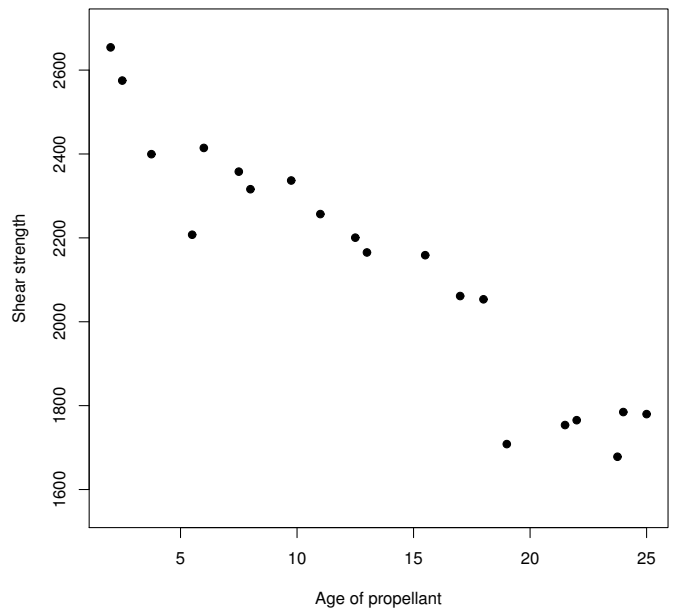
- **Background**: A rocket motor is manufactured by bonding an igniter propellant and a sustainer propellant together inside a mental housing. It is suspected that shear strength is related to the age in weeks of the batch of sustainer propellant. Twenty observations on shear strength and the age of the corresponding batch of propellant are shown in table example 1.2.
- y: the shear strength of the bond

  x: the age of propellant

Example 1.2: The Rocket Propellant Data

- **Question 1.2.1**: Plot the scatter diagram of shear strength versus propellant age.

```
1  data2 <- read.csv("example1.2.csv")
2  names(data2) <- c("Obs_i", "shear_strength_yi", "age_of_
       propellant_xi")                          #Rename the header
3
4  plot(data2$age_of_propellant_xi, data2$shear_strength_yi, xlab
       ="Age of propellant", ylab="Shear strength",pch=19,ylim=c
       (1555,2700))
```



Example 1.2: The Rocket Propellant Data

- **Question 1.2.2**: Fit a simple linear regression model relating shear strength y to the age of propellant x.

Method 1

```
x <- data2$age_of_propellant_xi
y <- data2$shear_strength_yi
xbar <- mean(x)                                  #obtain the mean
ybar <- mean(y)
Sxx <- sum((x-xbar)^2)                          #Obtain Sxx and Sxy
Sxy <- sum(y*(x-xbar))
betahat1 <- Sxy/Sxx                    #Obtain betahat1 and betahat2
betahat0 <- ybar-betahat1*xbar
```

- $\bar{y} = \frac{1}{n}\sum_{i=1}^{n} y_i = 2131.3575$ and $\bar{x} = \frac{1}{n}\sum_{i=1}^{n} x_i = -37.15$
- $S_{xx} = \sum_{i=1}^{n}(x_i - \bar{x})^2 = 1106.56$ and $S_{xy} = \sum_{i=1}^{n} y_i(x_i - \bar{x}) = -41,112.6$
- $\hat{\beta}_1 = \frac{S_{xy}}{S_{xx}} = -37.15$ and $\hat{\beta}_0 = \bar{y} - \hat{\beta}_1\bar{x} = 2627.82$
- The **least-square fit** is $\hat{y} = 2627.82 - 37.15x$

Example 1.2: The Rocket Propellant Data

Method 2

```
1  x <- data2$age_of_propellant_xi
2  y <- data2$shear_strength_yi
3  fit <- lm(y~x)
4  summary(fit)
```

```
Call:
lm(formula = y ~ x)

Residuals:
    Min      1Q  Median      3Q     Max
-215.98  -50.68   28.74   66.61  106.76

Coefficients:
            Estimate Std. Error t value Pr(>|t|)
(Intercept) 2627.822     44.184   59.48  < 2e-16 ***
x            -37.154      2.889  -12.86 1.64e-10 ***
---
Signif.codes: 0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 96.11 on 18 degrees of freedom
Multiple R-squared:  0.9018,    Adjusted R-squared:  0.8964
F-statistic: 165.4 on 1 and 18 DF,  p-value: 1.643e-10
```

- The **least-square fit** is $\hat{y} = 2627.82 - 37.15x$
- We may interpret the slope -37.15 as the average weekly decrease in propellant shear strength due to the age of the propellant. Since the lower limit of the x is near the origin, the intercept $2,627.82$ represents the shear strength in a batch of propellant immediately following manufacture.
- Several questions after obtaining the least-squares fit:
  1. How well does this equation fit the data?
  2. Is the model likely to be useful as a predictor?
  3. Are any of the basic assumptions (such as constant variance and uncorrelated errors) violated, and if so, how serious is this?

  $\Rightarrow$ **Residuals**:

  1. help evaluate model adequacy;
  2. can be reviewed as realizations of the model errors ε_is.

Data, Fitted Values, and Residuals for Example 1.2

	y_i	$\hat{y}$	e_i
1	2158.70	2051.94	106.76
2	1678.15	1745.42	-67.27
3	2316.00	2330.59	-14.59
4	2061.30	1996.21	65.09
5	2207.50	2423.48	-215.98
6	1708.30	1921.90	-213.60
7	1784.70	1736.14	48.56
8	2575.00	2534.94	40.06
9	2357.90	2349.17	8.73
10	2256.70	2219.13	37.57
11	2165.20	2144.83	20.37
12	2399.55	2488.50	-88.95
13	1779.80	1698.98	80.82
14	2336.75	2265.57	71.18
15	1765.30	1810.44	-45.14
16	2053.50	1959.06	94.44
17	2414.40	2404.90	9.50
18	2200.50	2163.40	37.10
19	2654.20	2553.52	100.68
20	1753.70	1829.02	-75.32
	$\sum y_i = 42,627.15$	$\sum \hat{y}_i = 42,627.15$	$\sum e_i = 0.00$

- Table above displays **the observed values** y_i, **the fitted values** $\hat{y}_i$, and **the residuals**.

R Regression Output for Example 1.2

	Dependent variable:
	y
x	-37.154^{***}
	(2.889)
Constant	$2{,}627.822^{***}$
	(44.184)
Observations	20
R^2	0.902
Adjusted R^2	0.896
Residual Std. Error	96.106 (df = 18)
F-Statistic	165.377^{***} (df = 1; 18)
Note:	*p<0.1; **p<0.05; ***p<0.01

Computer Output: Computer software packages are used extensively in fitting regression models. Regression routines are found in both network and PC-based statistical software, as well as in many popular spreadsheet packages.

Properties of the Least-squares Estimators

- The least-square estimators $\hat{\beta}_0$ and $\hat{\beta}_1$ are **linear combinations** of the observations y_i.

$$\hat{\beta}_1 = \frac{S_{xy}}{S_{xx}} = \frac{\sum_{i=1}^{n}(x_i - \bar{x})y_i}{S_{xx}} \equiv \sum_{i=1}^{n} c_i y_i,$$

where $c_i = (x_i - \bar{x})/S_{xx}$ for $i = 1, 2, \ldots, n$. And

$$\hat{\beta}_0 = \bar{y} - \hat{\beta}_1 \bar{x} = \frac{1}{n}\sum_{i=1}^{n} y_i - \bar{x}\sum_{i=1}^{n} c_i y_i = \sum_{i=1}^{n}(\frac{1}{n} - \bar{x}c_i)y_i$$

N.B.

  - $\sum_{i=1}^{n} c_i = 0$ and $\sum_{i=1}^{n} c_i x_i = 1$, and $\sum_{i=1}^{n} c_i^2 = \frac{\sum_{i=1}^{n}(x-\bar{x})^2}{S_{xx}^2} = \frac{1}{S_{xx}}$.

- $\hat{\beta}_1$ and $\hat{\beta}_0$ are **unbiased estimators**.

$$E(\hat{\beta}_1) = E(\sum_{i=1}^{n} c_i y_i) = \beta_0 \sum_{i=1}^{n} c_i + \beta_1 \sum_{i=1}^{n} c_i(x_i - \bar{x}) + \beta_1 \bar{x}\sum_{i=1}^{n} c_i = \beta_1 \frac{\sum_{i=1}^{n}(x_i - \bar{x})^2}{\sum_{i=1}^{n}(x_i - \bar{x})^2} = \beta_1$$

$$E(\hat{\beta}_0) = \sum_{i=1}^{n}(\frac{1}{n} - \bar{x}c_i)E(y_i) = \sum_{i=1}^{n}(\frac{1}{n} - \bar{x}c_i)(\beta_0 + \beta_1 x_i) = \beta_0 + \beta_1 \times \frac{1}{n}\sum_{i=1}^{n} x_i - \beta_1 \bar{x} = \beta_0$$

The expected value of the estimator $\hat{\beta}_1$ ($\hat{\beta}_0$) is equal to the true value of the parameter that it estimates, β_1 (β_0).

Properties of the Least-squares Estimators

- The (conditional) **variance** of $\hat{\beta}_1$ and $\hat{\beta}_0$ are

$$\mathrm{Var}(\hat{\beta}_1) = \mathrm{Var}(\sum_{i=1}^{n} c_i y_i) = \sum_{i=1}^{n} {c_i}^2 \mathrm{Var}(y_i) = \sigma^2 \sum_{i=1}^{n} {c_i}^2 = \frac{\sigma^2}{S_{xx}}$$

$$\mathrm{Var}(\hat{\beta}_0) = \sum_{i=1}^{n} (\frac{1}{n} - \bar{x}c_i)^2 \sigma^2 = \sigma^2 (\frac{1}{n} - 2\frac{\bar{x}}{n} \sum_{i=1}^{n} c_i + \bar{x}^2 \sum_{i=1}^{n} {c_i}^2) = \sigma^2 \left(\frac{1}{n} + \frac{\bar{x}^2}{S_{xx}} \right)$$

- **Gauss-Markov theorem**

  For the regression model with the assumptions $E(\varepsilon_i) = 0$, $\mathrm{Var}(\varepsilon_i) = \sigma^2$ and uncorrelated errors, the least-squares estimators are **unbiased** and have **minimum variance** when compared with all other unbiased estimators that are linear combinations of the y_i.

- **BLUE**: **best linear unbiased estimators**, where "best" implies minimum variance when compared with all other unbiased estimators that are linear combinations of the y_i.

Residuals

The ith **residual** is defined as

$$e_i = y_i - \hat{y}_i, \ \ i = 1, \ldots, n.$$

1. $\sum(y_i - \hat{y}_i) = \sum e_i = 0$;
2. $\sum y_i = \sum \hat{y}_i$;
3. The least-squares regression line always passes through the **centroid** [the point $(\bar{x}, \bar{y})$] of the data;
4. $\sum x_i e_i = 0$; (Hint: by the second normal equation.)
5. $\sum \hat{y}_i e_i = 0$. (Hint: by the 1st and the previous property.)

Target 2: Estimation of σ^2

- n pairs of sample data: $(y_i, x_i), \quad i = 1, \ldots, n$
- **Sample SLR model**: $y_i = \beta_0 + \beta_1 x_i + \varepsilon_i, \quad i = 1, \ldots, n$
- **Assumptions**: Assumptions 1, 2 & 3
- **Residual** (or **error**) **sum of squares**: $SS_{Res} = \sum_{i=1}^{n} e_i^2 = \sum_{i=1}^{n} (y_i - \hat{y}_i)^2$.
- Since $SS_{Res}(\beta_0, \beta_1) \sim \sigma^2 \chi^2_{n-2}$ and $E(SS_{Res}) = (n-2)\sigma^2$,

$$E(\hat{\sigma}^2) = E(\frac{SS_{Res}}{n-2}) = \frac{E(SS_{Res})}{n-2} = \sigma^2$$

An **unbiased** estimator of σ^2 is

$$\hat{\sigma}^2 = \frac{SS_{Res}}{n-2} = MS_{Res}$$

where MS_{Res} is called the **residual mean square**. The square root of $\hat{\sigma}^2$ is called the **standard error of regression**.

- Remark: Since $\hat{\sigma}^2$ is computed from the regression model residuals, we say that it is a **model-dependent** estimate of σ^2.

Example 1.2: The Rocket Propellant Data

- **Question 1.2.3**: Estimate the error variance σ^2.

```
1    anova(fit)

     Analysis of Variance Table

     Response: y
               Df  Sum Sq Mean Sq F value    Pr(>F)
     x          1 1527483 1527483  165.38 1.643e-10 ***
     Residuals 18  166255    9236
     ---
     Signif.codes: 0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

$$\hat{\sigma}^2 = \frac{SS_{Res}}{n-2} = \frac{166255}{18} = 9236.3889$$

Alternate Form of the Model

- Redefine the regressor variable $x_i \rightarrow x_i - \bar{x}$, then

$$y_i = \beta_0' + \beta_1 (x_i - \bar{x}) + \varepsilon_i \quad i = 1, \ldots, n.$$

  where $\beta_0' = \beta_0 + \beta_1 \bar{x}$.

- $\hat{\beta}_0' = \bar{y}$ and $\hat{\beta}_1 = \frac{S_{yx}}{S_{xx}}$
- **Advantage**:
  - The least-squares estimators $\hat{\beta}_0' = \bar{y}$ and $\hat{\beta}_1 = \frac{S_{yx}}{S_{xx}}$ are **uncorrelated**, that is $\text{Cov}(\hat{\beta}_0', \hat{\beta}_1) = 0$;
  - This will make some applications of the model easier, such as finding confidence intervals on the mean of y;
  - The fitted model is $\hat{y} = \bar{y} + \hat{\beta}_1 (x - \bar{x})$ directly reminds the analyst that the regression model is only valid over the range of x in the **original data**. This region is centered at $\bar{x}$.

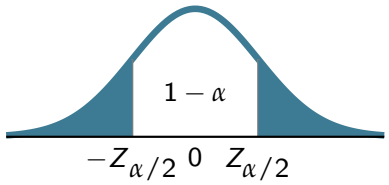
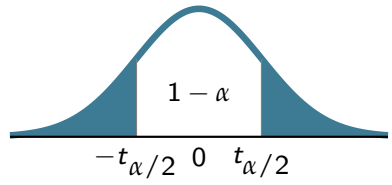
Target 3: Hypothesis Testing on β_0 and β_1

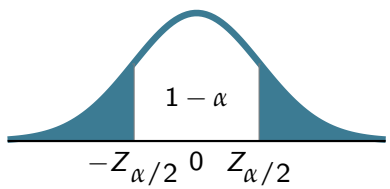
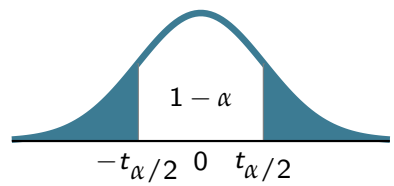
- n pairs of sample data: $(y_i, x_i), \quad i = 1, \ldots, n$
- **Sample SLR model**: $y_i = \beta_0 + \beta_1 x_i + \varepsilon_i, \quad i = 1, \ldots, n$
- **Assumption**: $\varepsilon_1, \varepsilon_2, \ldots, \varepsilon_n \overset{\text{i.i.d.}}{\sim} N(0, \sigma^2)$
- **Hypothesis**: $\begin{cases} \text{Q1}: H_0: \beta_1 = \beta_{10} \text{ vs. } H_1: \beta_1 \neq \beta_{10} \\ \text{Q2}: H_0: \beta_0 = \beta_{00} \text{ vs. } H_1: \beta_0 \neq \beta_{00} \end{cases}$
- For Q1: $\boldsymbol{H_0: \beta_1 = \beta_{10} \text{ vs. } H_1: \beta_1 \neq \beta_{10}}$ α: significance level
  - If σ^2 is **known**
    - $\hat{\beta}_1 \sim N\left(\beta_1, \frac{\sigma^2}{S_{xx}}\right)$
    - Under H_0:
    $$Z_0 = \frac{\hat{\beta}_1 - \beta_{10}}{\sqrt{\sigma^2 / S_{xx}}} \sim N(0,1)$$
    - Reject H_0 in favor of H_1 if $|Z_0| > Z_{\alpha/2}$.
  - If σ^2 is **unknown**
    - $(n-2)MS_{Res}/\sigma^2 \sim \chi^2_{n-2}$ and $MS_{Res} \perp \hat{\beta}_1$
    - Under H_0:
    $$t_0 = \frac{\hat{\beta}_1 - \beta_{10}}{\sqrt{MS_{Res}/S_{xx}}} \equiv \frac{\hat{\beta}_1 - \beta_{10}}{\text{se}(\hat{\beta}_1)} \sim t_{n-2}$$
    - Reject H_0 in favor of H_1 if $|t_0| > t_{\alpha/2, n-2}$.

Remark: $\text{se}(\hat{\beta}_1)$ is called the **(estimated) standard error** of the slope.

- For Q2: $H_0 : \beta_0 = \beta_{00}$ **vs.** $H_1 : \beta_0 \neq \beta_{00}$ α: significance level
  - If σ^2 is **known**
    - $\hat{\beta}_0 \sim N\left(\beta_0, \sigma^2\left(\frac{1}{n} + \frac{\bar{x}^2}{S_{xx}}\right)\right)$
    - Under H_0:
$$Z_0 = \frac{\hat{\beta}_0 - \beta_{00}}{\sqrt{\sigma^2\left(\frac{1}{n} + \frac{\bar{x}^2}{S_{xx}}\right)}} \sim N(0,1)$$
    - Reject H_0 in favor of H_1 if $|Z_0| > Z_{\alpha/2}$.
  - If σ^2 is **unknown**
    - $(n-2)MS_{Res}/\sigma^2 \sim \chi^2_{n-2}$, and $MS_{Res} \perp \hat{\beta}_0$
    - Under H_0:
$$t_0 = \frac{\hat{\beta}_0 - \beta_{00}}{\sqrt{MS_{Res}\left(\frac{1}{n} + \frac{\bar{x}^2}{S_{xx}}\right)}} \equiv \frac{\hat{\beta}_0 - \beta_{00}}{\text{se}(\hat{\beta}_0)} \sim t_{n-2}$$
    - Reject H_0 in favor of H_1 if $|t_0| > t_{\alpha/2,n-2}$.

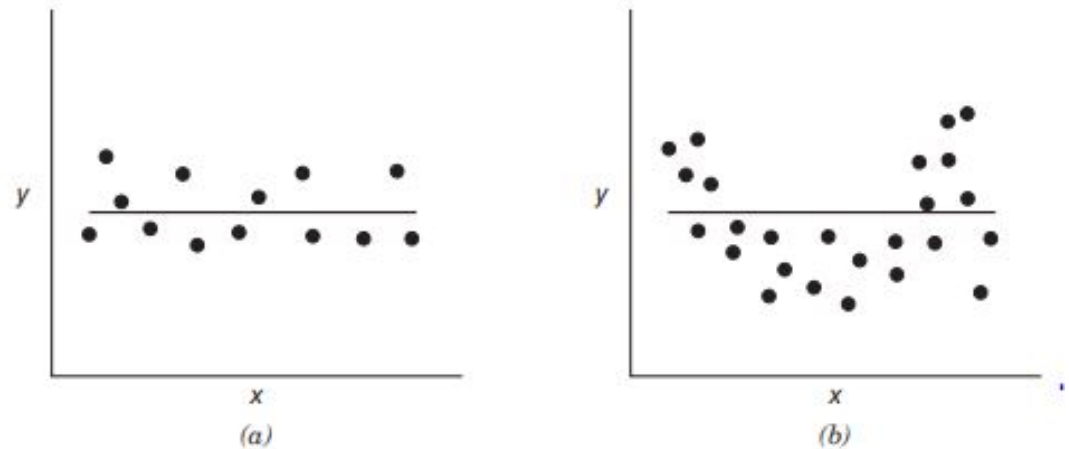
  Remark: $\text{se}(\hat{\beta}_0)$ called the **(estimated) standard error** of the intercept.
- **P-value approach** could also be used for decision making.

Hypothesis	$H_0 : \beta_1 = \beta_{10}$ vs. $H_1 : \beta_1 \neq \beta_{10}$	
σ^2	known	unknown
Condition	$\hat{\beta}_1 \sim N\left(\beta_1, \frac{\sigma^2}{S_{xx}}\right)$	$(n-2)MS_{Res}/\sigma^2 \sim \chi^2_{n-2}$ and $MS_{Res} \perp \hat{\beta}_1$
Test Statistic	$Z_0 = \frac{\hat{\beta}_1 - \beta_{10}}{\sqrt{\sigma^2/S_{xx}}} \sim N(0,1)$	$t_0 = \frac{\hat{\beta}_1 - \beta_{10}}{\sqrt{MS_{Res}/S_{xx}}} \equiv \frac{\hat{\beta}_1 - \beta_{10}}{\text{se}(\hat{\beta}_1)} \sim t_{n-2}$
Graph		
Rejection Region	$\lvert Z_0 \rvert > Z_{\alpha/2}$	$\lvert t_0 \rvert > t_{\alpha/2,n-2}$

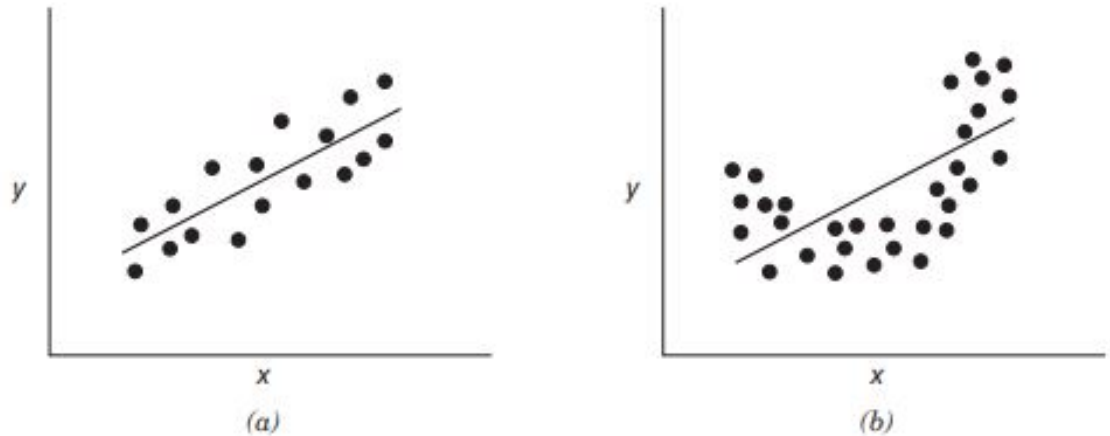
Hypothesis	$H_0 : \beta_0 = \beta_{00}$ vs. $H_1 : \beta_0 \neq \beta_{00}$	
σ^2	known	unknown
Condition	$\hat{\beta}_0 \sim N\left(\beta_0, \sigma^2\left(\frac{1}{n} + \frac{\bar{x}^2}{S_{xx}}\right)\right)$	$(n-2)MS_{Res}/\sigma^2 \sim \chi^2_{n-2}$ and $MS_{Res} \perp \hat{\beta}_1$
Test Statistic	$Z_0 = \frac{\hat{\beta}_0 - \beta_{00}}{\sqrt{\sigma^2\left(\frac{1}{n} + \frac{\bar{x}^2}{S_{xx}}\right)}} \sim N(0,1)$	$t_0 = \frac{\hat{\beta}_0 - \beta_{00}}{\sqrt{MS_{Res}\left(\frac{1}{n} + \frac{\bar{x}^2}{S_{xx}}\right)}} \equiv \frac{\hat{\beta}_0 - \beta_{00}}{\text{se}(\hat{\beta}_0)} \sim t_{n-2}$
Graph		
Rejection Region	$\lvert Z_0 \rvert > Z_{\alpha/2}$	$\lvert t_0 \rvert > t_{\alpha/2,n-2}$

Testing Significance of Regression

- Hypothesis: $H_0 : \beta_1 = 0$ **vs.** $H_1 : \beta_1 \neq 0$ α: significance level
- **Fail to reject** H_0: there is **no linear relationship** between y and x.
  - x is of little value in explaining the variation in y and that the best estimator of y for any x is $\hat{y} = \bar{y}$ (Figure (a));
  - The true relationship between x and y is not linear (Figure (b)).



- **Reject H_0**: x is **of value in explaining the variability** in y.
  - The straight-line model is adequate (Figure (a));
  - Even though there is a linear effect of x, better results could be obtained with **the addition of higher order polynomial terms** in x (Figure (b)).



- The test procedure has two approaches:
  - t test: Under H_0:

$$t_0 = \frac{\hat{\beta}_1}{\text{se}(\hat{\beta}_1)}$$

    Reject H_0 in favor of H_1 if $|t_0| > t_{\alpha/2,n-2}$.
  - Analysis-of-variance (**ANOVA**)

Analysis-of-variance (ANOVA)

- **Fundamental analysis-of-variance identity for a regression model**:

$$\sum_{i=1}^{n}(y_i - \bar{y})^2 = \sum_{i=1}^{n}(\hat{y}_i - \bar{y}_i)^2 + \sum_{i=1}^{n}(y_i - \hat{y}_i)^2$$

$$SS_T = SS_R + SS_{Res}$$

where

- $SS_T = \sum_{i=1}^{n}(y_i - \bar{y})^2$
  total sum of squares
  (variation around mean);
- $SS_R = \sum_{i=1}^{n}(\hat{y}_i - \bar{y}_i)^2 = \hat{\beta}_1 S_{xy}$
  regression/model sum of squares
  (variation between regression line and mean);
- $SS_{Res} = \sum_{i=1}^{n}(y_i - \hat{y}_i)^2$
  residual/error sum of squares
  (variation around regression line).

- **Degree-of-freedom** (df): the number of **independent values** that can vary in an analysis **without breaking any constraints**.

$$df_T = df_R + df_{Res}$$

$$n - 1 = 1 + (n - 2)$$

where SS_T: $df_T = n - 1$, SS_R: $df_R = 1$, and SS_{Res}: $df_{Res} = n - 2$.

  - $\boldsymbol{df_T = n - 1}$ because one degree of freedom is lost as a result of the constraint $\sum_{i=1}^{n} (y_i - \bar{y})$ on the deviations $y_i - \bar{y}$.
  - $\boldsymbol{df_R = 1}$ because $SS_R = \hat{\beta}_1 S_{xy}$ and SS_R is completely determined by one parameter, namely, $\hat{\beta}_1$.
  - $\boldsymbol{df_{Res} = n - 2}$ because two constraints are imposed on the deviations $y_i - \hat{y}_i$ as a result of estimating $\hat{\beta}_0$ and $\hat{\beta}_1$.

ANOVA for Testing Significance of Regression

Source of Variation	Sum of Squares	Degrees of Freedom	Mean Square	F_0
Regression	SS_R	1	$MS_R = SS_R/1$	MS_R/MS_{Res}
Residuals	SS_{Res}	$n-2$	$MS_{Res} = SS_{Res}/(n-2)$	
Total	SS_T	$n-1$		

- Testing significance of regression, then ***F*-test** will be used.
- H_0: $\beta_1 = 0$ vs. H_1 : $\beta_1 \neq 0$ α: significance level.
- $SS_{Res}/\sigma^2 = (n-2)MS_{Res}/\sigma^2 \sim \chi^2_{n-2}$, $SS_R/\sigma^2 = MS_R/\sigma^2 \sim \chi^2_1$, and $SS_{Res} \perp SS_R$.
- Under H_0:

$$F_0 = \frac{SS_R/1}{SS_{Res}/(n-2)} = MS_R/MS_{Res} \sim F_{1,n-2}$$

- Reject H_0 in favor of H_1 if $F_0 > F_{\alpha,1,n-2}$.

- **Remark**:
  - $E(MS_{Res}) = \sigma^2$, $E(MS_R) = \sigma^2 + \beta_1^2 S_{xx}$. This expected mean squares indicate that if the observed value of F_0 is large, then it is likely that the slope $\beta_1 \neq 0$;
  - If $\beta_1 \neq 0$, then F_0 follows a noncentral F distribution with 1 and $n-2$ degrees of freedom and a non-centrality parameter of $\lambda = \frac{\beta_1^2 S_{xx}}{\sigma^2}$, which alos indicates that the observed value of F_0 should be large if $\beta_1 \neq 0$.

Example 1.2: The Rocket Propellant Data

- **Question 1.2.4**: Test for significance of regression at $\alpha = 0.05$.

Method 1 (t-test)

```
1   summary(fit)
```

```
Call:
lm(formula = y ~ x)

Residuals:
    Min      1Q  Median      3Q     Max
-215.98  -50.68   28.74   66.61  106.76

Coefficients:
            Estimate Std. Error t value Pr(>|t|)
(Intercept) 2627.822     44.184   59.48  < 2e-16 ***
x            -37.154      2.889  -12.86 1.64e-10 ***
---
Signif.codes: 0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 96.11 on 18 degrees of freedom
Multiple R-squared:  0.9018,    Adjusted R-squared:  0.8964
F-statistic: 165.4 on 1 and 18 DF,  p-value: 1.643e-10
```

Example 1.2: The Rocket Propellant Data

- The estimate of the slope is $\hat{\beta}_1 = -37.15$, and in Question 1.2.3, we computed the estimate of σ^2 to be 9236.3889. The standard error of the slope is

$$se(\hat{\beta}_1) = \sqrt{\frac{MS_{Res}}{S_{xx}}} = \sqrt{\frac{9236.3889}{1106.56}} = 2.889$$

- Therefore, the **test statistic** is

$$t_0 = \frac{\hat{\beta}_1}{se(\hat{\beta}_1)} = \frac{-37.15}{2.89} = -12.86$$

- If we choose $\alpha = 0.05$, the **critical value** of t is $t_{0.025,18} = 2.101$. Thus, we are against $H_0 : \beta_1 = 0$ in favor of H_1 and conclude that there is a linear relationship between shear strength and the age of the propellant.

Example 1.2: The Rocket Propellant Data

Method 2 (ANOVA)

```
1    anova(fit)
```

```
Analysis of Variance Table

Response: y
Df  Sum Sq Mean Sq F value    Pr(>F)
x          1 1527483 1527483  165.38 1.643e-10 ***
Residuals 18  166255    9236
---
Signif.codes: 0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

- $F_0 = 165.38$ and from F-distribution table, $F_{0.01,1,18} = 8.29$. The P-value for this test is 1.643×10^{-10}.
- Consequently, we reject $H_0 : \beta_1 = 0$ in favor of H_1 and conclude that there is a linear relationship between shear strength and the age of the propellant.

More about the t-test

$$t_0^2 = \frac{\hat{\beta}_1^2 S_{xx}}{MS_{Res}} = \frac{\hat{\beta}_1 S_{xy}}{MS_{Res}} = \frac{MS_R}{MS_{Res}} = F_0$$

- In general, the square of a t random variable with f degrees of freedom is an F random variable with one and f degrees of freedom in the numerator and denominator, respectively.
- Although **the t test for $H_0 : \beta_1 = 0$ is equivalent to the F test in SLR**, the t test is somewhat more adaptable, as it could be used for **one-sided alternative hypotheses** (either $H_1 : \beta_1 < 0$ or $H_1 : \beta_1 > 0$), while the F test considers only the **two-sided alternative hypothesis**.
- Deciding that $\beta_1 = 0$ is a very important conclusion that is only aided by the t or F test.
- Remark: The inability to show that the slope is not statistically different from zero may not necessarily mean that y and x are unrelated. It may mean that our ability to detect this relationship has been obscured by the variance of the measurement process or that the range of values of x is inappropriate.

Target 4: Confidence Interval on β_0 and β_1

- n pairs of sample data: $(y_i, x_i), \quad i = 1, \ldots, n$
- **Sample SLR model**: $y_i = \beta_0 + \beta_1 x_i + \varepsilon_i, \qquad i = 1, \ldots, n$
- **Assumptions**: $\varepsilon_1, \varepsilon_2, \ldots, \varepsilon_n \overset{\text{i.i.d.}}{\sim} N(0, \sigma^2)$
- $E(\hat{\beta}_0) = \beta_0 \quad$ and $\quad \mathrm{Var}(\hat{\beta}_1) = \frac{\sigma^2}{S_{xx}}$;

  $E(\hat{\beta}_1) = \beta_1 \quad$ and $\quad \mathrm{Var}(\hat{\beta}_0) = \sigma^2 \left(\frac{1}{n} + \frac{\bar{x}^2}{S_{xx}} \right)$
- Since the variances of $\hat{\beta}_0$ and $\hat{\beta}_1$ involve the **unknown parameter** σ^2, **additional results** $SS_{Res}/\sigma^2 \sim \chi^2_{n-2}$ and $SS_{Res} \perp (\hat{\beta}_0, \hat{\beta}_1)$ are needed. For β_1,

$$\frac{\hat{\beta}_1 - \beta_1}{\sqrt{\hat{\sigma}^2 / S_{xx}}} = \frac{\hat{\beta}_1 - \beta_1}{\sqrt{\sigma^2 / S_{xx}}} \Bigg/ \sqrt{\frac{(n-2)\hat{\sigma}^2}{(n-2)\sigma^2}} = \frac{\hat{\beta}_1 - \beta_1}{\sqrt{\sigma^2 / S_{xx}}} \Bigg/ \sqrt{\frac{SS_{Res}}{(n-2)\sigma^2}} \sim t_{n-2}$$

$$P\left(-t_{\alpha/2, n-2} \le \frac{\hat{\beta}_1 - \beta_1}{\sqrt{\hat{\sigma}^2 / S_{xx}}} \le t_{\alpha/2, n-2} \right) = 1 - \alpha$$

$$\Rightarrow \quad P\left(\hat{\beta}_1 - t_{\alpha/2, n-2} \sqrt{\frac{\hat{\sigma}^2}{S_{xx}}} \le \beta_1 \le \hat{\beta}_1 + t_{\alpha/2, n-2} \sqrt{\frac{\hat{\sigma}^2}{S_{xx}}} \right) = 1 - \alpha$$

- Therefore, a $100(1-\alpha)$ percent **confidence interval** (CI) on β_1 is

$$\hat{\beta}_1 - t_{\alpha/2,n-2}\text{se}(\hat{\beta}_1) \leq \beta_1 \leq \hat{\beta}_1 + t_{\alpha/2,n-2}\text{se}(\hat{\beta}_1)$$

- Similarly, a $100(1-\alpha)$ percent **confidence interval** (CI) on β_0 is

$$\hat{\beta}_0 - t_{\alpha/2,n-2}\text{se}(\hat{\beta}_0) \leq \beta_0 \leq \hat{\beta}_0 + t_{\alpha/2,n-2}\text{se}(\hat{\beta}_0)$$

- Remark: $\text{se}(\hat{\beta}_1) = \sqrt{\frac{\hat{\sigma}^2}{S_{xx}}}$

$$\text{se}(\hat{\beta}_0) = \sqrt{\hat{\sigma}^2\left(\frac{1}{n} + \frac{\bar{x}^2}{S_{xx}}\right)}$$

Confidence Interval on σ^2

- From the distribution $SS_{Res}/\sigma^2 \sim \chi^2_{n-2}$,

$$P\left(\chi^2_{1-\alpha/2,n-2} \leq \frac{SS_{Res}}{\sigma^2} \leq \chi^2_{\alpha/2,n-2}\right) = 1-\alpha$$

$$\Rightarrow \quad P\left(\frac{(n-2)\hat{\sigma}^2}{\chi^2_{\alpha/2,n-2}} \leq \sigma^2 \leq \frac{(n-2)\hat{\sigma}^2}{\chi^2_{1-\alpha/2,n-2}}\right) = 1-\alpha$$

- Therefore, a $100(1-\alpha)$ percent **confidence interval** (CI) on σ^2 is

$$\left(\frac{(n-2)MS_{Res}}{\chi^2_{\alpha/2,n-2}}, \frac{(n-2)MS_{Res}}{\chi^2_{1-\alpha/2,n-2}}\right)$$

- Remark: $\hat{\sigma}^2 = \frac{SS_{Res}}{n-2} = MS_{Res}$

Summary of CI on β_0, β_1 and σ^2

- **Assumptions**: $\varepsilon_1, \varepsilon_2, \ldots, \varepsilon_n \overset{\text{i.i.d.}}{\sim} N(0, \sigma^2)$
- $\frac{\hat{\beta}_1 - \beta_1}{\text{se}(\hat{\beta}_1)} \sim t_{n-2}$ and $\frac{\hat{\beta}_0 - \beta_0}{\text{se}(\hat{\beta}_0)} \sim t_{n-2}$
- A $100(1-\alpha)$ percent **confidence interval (CI)** on the slope β_0 is given by

$$\hat{\beta}_0 - t_{\alpha/2,n-2}\text{se}(\hat{\beta}_0) \leq \beta_0 \leq \hat{\beta}_0 + t_{\alpha/2,n-2}\text{se}(\hat{\beta}_0)$$

- A $100(1-\alpha)$ percent **confidence interval (CI)** on the intercept β_1 is given by

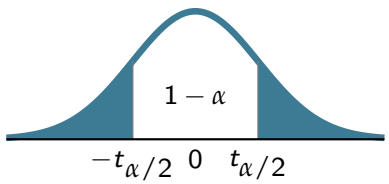
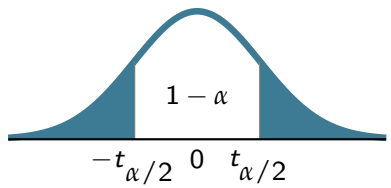
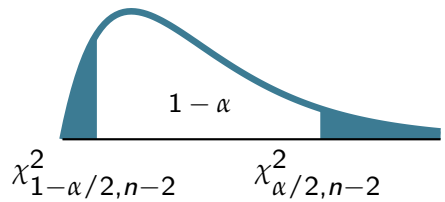
$$\hat{\beta}_1 - t_{\alpha/2,n-2}\text{se}(\hat{\beta}_1) \leq \beta_1 \leq \hat{\beta}_1 + t_{\alpha/2,n-2}\text{se}(\hat{\beta}_1)$$

- A $100(1-\alpha)$ percent **confidence interval (CI)** on σ^2 is given by

$$\frac{(n-2)MS_{Res}}{\chi^2_{\alpha/2,n-2}} \leq \sigma^2 \leq \frac{(n-2)MS_{Res}}{\chi^2_{1-\alpha/2,n-2}}$$

- Remarks: $\chi^2_{1-\alpha/2,n-2}$: lower $1-\alpha/2$ quantile for χ^2-distribution

  $\chi^2_{\alpha/2,n-2}$: upper $\alpha/2$ quantile for χ^2-distribution

Assumptions	$\varepsilon_1, \varepsilon_2, \ldots, \varepsilon_n \overset{\text{i.i.d.}}{\sim} N(0, \sigma^2)$		
Parameter	β_0	β_1	σ^2
Distribution	$\hat{\beta}_1 \sim N\left(\beta_1, \frac{\sigma^2}{S_{xx}}\right)$	$\hat{\beta}_0 \sim N\left(\beta_0, \sigma^2\left(\frac{1}{n} + \frac{\bar{x}^2}{S_{xx}}\right)\right)$	$\frac{SS_{Res}}{\sigma^2} \sim \chi^2_{n-2}$
Confident Level	$P\Big(-t_{\alpha/2,n-2} \leq \frac{\hat{\beta}_1 - \beta_1}{\text{se}(\hat{\beta}_1)} \leq t_{\alpha/2,n-2}\Big) = 1-\alpha$	$P\Big(-t_{\alpha/2,n-2} \leq \frac{\hat{\beta}_0 - \beta_0}{\text{se}(\hat{\beta}_0)} \leq t_{\alpha/2,n-2}\Big) = 1-\alpha$	$P\Big(\chi^2_{1-\alpha/2,n-2} \leq \frac{SS_{Res}}{\sigma^2} \leq \chi^2_{\alpha/2,n-2}\Big) = 1-\alpha$
Graph			
Confident Interval	$(\hat{\beta}_0 - t_{\alpha/2,n-2}\text{se}(\hat{\beta}_0), \hat{\beta}_0 + t_{\alpha/2,n-2}\text{se}(\hat{\beta}_0))$	$(\hat{\beta}_1 - t_{\alpha/2,n-2}\text{se}(\hat{\beta}_1), \hat{\beta}_1 + t_{\alpha/2,n-2}\text{se}(\hat{\beta}_1))$	$\left(\frac{(n-2)MS_{Res}}{\chi^2_{\alpha/2,n-2}}, \frac{(n-2)MS_{Res}}{\chi^2_{1-\alpha/2,n-2}}\right)$

Target 4: Interval Estimation of the Mean Response

- n pairs of sample data: $(y_i, x_i), \quad i = 1, \ldots, n$
- **Sample SLR model**: $y_i = \beta_0 + \beta_1 x_i + \varepsilon_i, \qquad i = 1, \ldots, n$
- **Assumptions**: $\varepsilon_1, \varepsilon_2, \ldots, \varepsilon_n \overset{\text{i.i.d.}}{\sim} N(0, \sigma^2)$
- $\widehat{E(y|x_0)} = \hat{\mu}_{y|x_0} = \hat{\beta}_0 + \hat{\beta}_1 x_0;$
- $\begin{aligned} \text{Var}(\hat{\mu}_{y|x_0}) &= \text{Var}(\hat{\beta}_0 + \hat{\beta}_1 x_0) = \text{Var}[\bar{y} + \hat{\beta}_1 (x_0 - \bar{x})] \\ &= \frac{\sigma^2}{n} + \frac{\sigma^2 (x_0 - \bar{x})^2}{S_{xx}} = \sigma^2 \left[\frac{1}{n} + \frac{(x_0 - \bar{x})^2}{S_{xx}} \right] \end{aligned}$
- $$\frac{\hat{\mu}_{y|x_0} - E(y|x_0)}{\sqrt{MS_{Res}\left(\frac{1}{n} + \frac{(x_0 - \bar{x})^2}{S_{xx}}\right)}} \sim t_{n-2}.$$
- A $100(1-\alpha)$ percent **confidence interval** (CI) on the **mean response** at the point $x = x_0$ is given by

$$\hat{\mu}_{y|x_0} - t_{\alpha/2 n-2} \sqrt{MS_{Res}\left(\frac{1}{n} + \frac{(x_0 - \bar{x})^2}{S_{xx}}\right)} \leq \text{E}(y|x_0) \leq \hat{\mu}_{y|x_0} + t_{\alpha/2, n-2} \sqrt{MS_{Res}\left(\frac{1}{n} + \frac{(x_0 - \bar{x})^2}{S_{xx}}\right)}$$

Example 1.2: The Rocket Propellant Data

- **Question 1.2.5**: Construct a 95% confidence interval for β_1.
  - The standard error of $\hat{\beta}_1$ is $se(\hat{\beta}_1) = 2.89$ and $t_{0.025,18} = 2.101$. Therefore, the 95% CI on the slope is

$$\hat{\beta}_1 - t_{0.025,18}\text{se}(\hat{\beta}_1) \leq \beta_1 \leq \hat{\beta}_1 + t_{0.025,18}\ \text{se}(\hat{\beta}_1)$$

$$-37.15 - (2.101)(2.89) \leq \beta_1 \leq -37.15 + (2.101)(2.89)$$

$$-43.22 \leq \beta_1 \leq -31.08$$

- **Question 1.2.6**: Find a 95% confidence interval for the expected shear strength if the age of propellant is 13.3625.

$$\hat{\mu}_{y|x_0} - t_{\alpha/2n-2}\sqrt{MS_{Res}\left(\frac{1}{n} + \frac{(x_0-\bar{x})^2}{S_{xx}}\right)} \leq \text{E}(y|x_0) \leq \hat{\mu}_{y|x_0} + t_{\alpha/2,n-2}\sqrt{MS_{Res}\left(\frac{1}{n} + \frac{(x_0-\bar{x})^2}{S_{xx}}\right)}$$

$$\hat{\mu}_{y|x_0} - (2.1)\sqrt{9236.39(\frac{1}{20} + \frac{(x_0-13.3625)^2}{1106.56})} \leq \text{E}(y|x_0) \leq \hat{\mu}_{y|x_0} + (2.1)\sqrt{9236.39(\frac{1}{20} + \frac{(x_0-13.3625)^2}{1106.56})}$$

If $x_0 = \bar{x} = 13.3625$, then $\hat{\mu}_{y|x_0} = 2131.40$, and the CI becomes $(2086.24, 2176.55)$.

Target 5: Prediction of New Observations

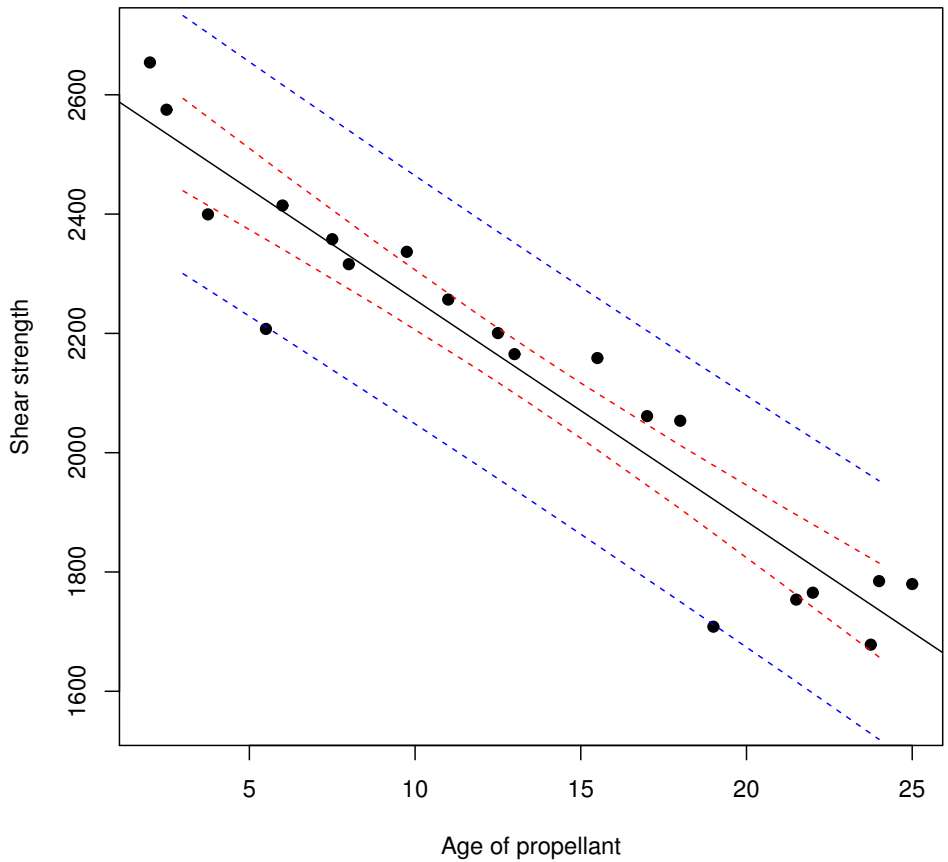
- n pairs of sample data: $(y_i, x_i), \quad i = 1, \ldots, n$
- **Sample SLR model**: $y_i = \beta_0 + \beta_1 x_i + \varepsilon_i, \qquad i = 1, \ldots, n$
- **Assumptions**: $\varepsilon_1, \varepsilon_2, \ldots, \varepsilon_n \overset{\text{i.i.d.}}{\sim} N(0, \sigma^2)$
- $\hat{y}_0 = \hat{\beta}_0 + \hat{\beta}_1 x_0$
- $\psi = y_0 - \hat{y}_0$, then

$$\mathrm{Var}(\psi) = \mathrm{Var}(y_0 - \hat{y}_0) = \mathrm{Var}(y_0) + \mathrm{Var}(\hat{y}_0) = \sigma^2 \left[1 + \frac{1}{n} + \frac{(x_0 - \bar{x})^2}{S_{xx}}\right]$$

- A $100(1-\alpha)$ percent **prediction interval** on a **future observation** at x_0 is given by

$$\hat{y}_0 - t_{\alpha/2,n-2}\sqrt{MS_{Res}\left(1 + \tfrac{1}{n} + \tfrac{(x_0-\bar{x})^2}{S_{xx}}\right)} \le y_0 \le \hat{y}_0 + t_{\alpha/2,n-2}\sqrt{MS_{Res}\left(1 + \tfrac{1}{n} + \tfrac{(x_0-\bar{x})^2}{S_{xx}}\right)}$$

Confidence Interval vs. Prediction Interval



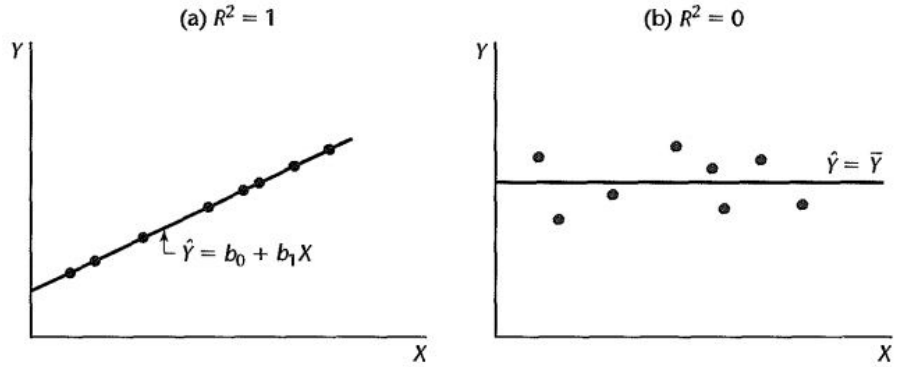
Red dashed lines: 95% **confidence interval**

Blue dashed lines: 95% **prediction interval**

Coefficient of Determination

$$R^2 = \frac{SS_R}{SS_T} = 1 - \frac{SS_{Res}}{SS_T}$$

- Remark:
  - R^2 is a **descriptive measure** that is frequently used in practice to describe the **degree of linear association between** x **and** y.
  - Since SS_T is a measure of the variability in y without considering the effect of the regressor variable x and SS_{Res} is a measure of the variability in y remaining after x has been considered, R^2 is often called **the proportion of variation explained** by the independent variable x.
  - Since $0 \leq SS_{Res} \leq SS_T$, it follows that $0 \leq R^2 \leq 1$.



- Values of R^2 that are **close to 1** imply that **most** of the variability in y is explained by the regression model.
- The **limiting values** of R^2 occur as follows:
  - When all observations fall on the fitted regression line, then SS_{Res} and $\underline{R^2 = 1}$. The predictor variable x accounts for all variation in the observations y_i. (figure(a))
  - When the fitted regression line is horizontal so that $\beta_1 = 0$ and $y_i = \bar{y}$, then $SS_{Res} = SS_T$ and $\underline{R^2 = 0}$. There is no linear association between x and y in the sample data, and the predictor variable x is of no help in reducing the variation in the observations y_i with linear regression. (figure(b))
- In practice, R^2 is not likely to be 0 or 1 but somewhere **between these limits**. Generally, R^2 cannot be exactly equal to 1 because the model cannot explain the variability related to "pure" error.

R^2 should be used with caution!

- It is always possible to make R^2 **large** by **adding enough terms to the model**.
- Although R^2 cannot decrease if we add a regressor variable to the model, this does not necessarily mean the new model is superior to the old one.
  - Unless the SS_{Res} in the new model is reduced by an amount equal to the original error mean square;
  - The new model will have a larger error mean square than the old one because of the loss of one degree of freedom for error. **(Recall:** $SS_{Res}/df_{Res} = MS_{Res}$**)**
- Generally R^2 will increase as the spread of the x's increases and decrease as the spread of the x's decreases provided the assumed model form is correct.

- The expected value of R^2 from a straight-line regression is approximately

$$E(R^2) \approx \frac{\beta_1^2 S_{xx}/(n-1)}{\frac{\beta_1^2 S_{xx}}{n-1} + \sigma^2}$$

  The expected value of R^2 will increase (decrease) as S_{xx} increases (decreases).
  - A large value of R^2 may result because x has been varied over an unrealistically large range.
  - R^2 may be small because the range of x was too small to allow its relationship with y to be detected.

- Some **misconceptions** about R^2:
  - A large value of R^2 implies a steep slope.

    Answer: **False**. The value of R^2 does not measure the magnitude of the slope of the regression line.
  - Large R^2 implies the appropriateness of the linear model.

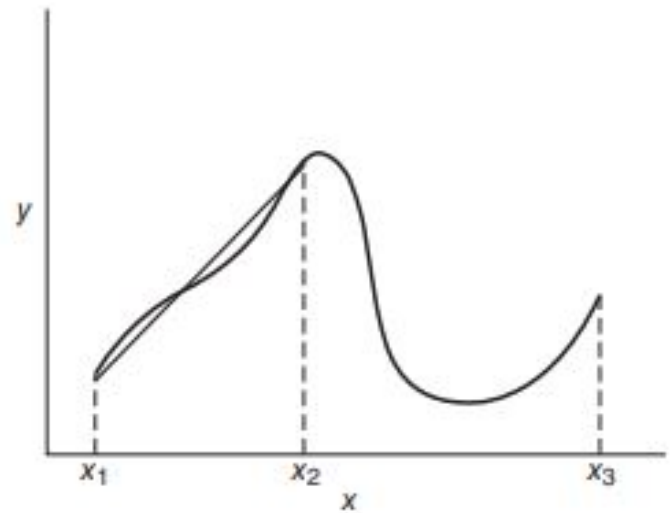
    Answer: **False**. Even y and x are nonlinearly related, R^2 will also be large.
  - Large R^2 implies that the regression model is an accurate predictor.

    Answer: **False**. Even the regression is not an accurate predictor, R^2 can still be large.

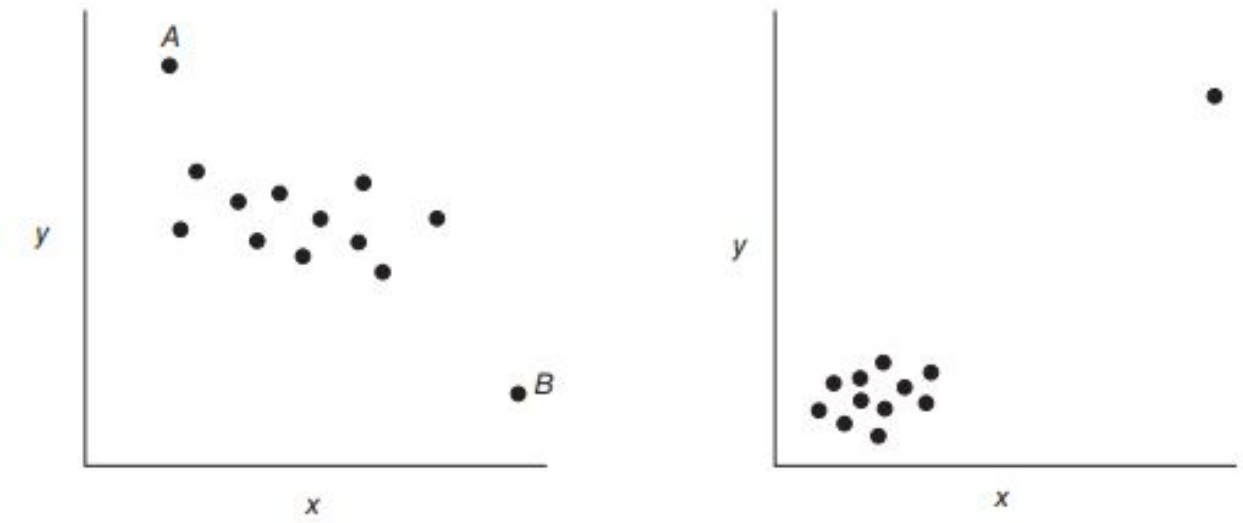
Some Consideration in the Use of Regression

There are several common abuses of regression that should be mentioned:

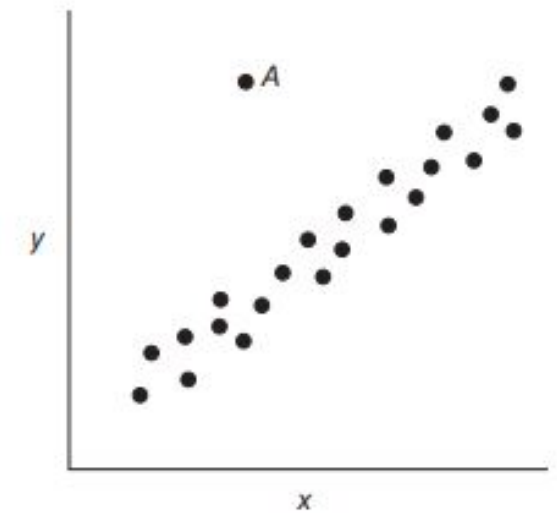
- **Abuse I**: Regression models are intended as interpolation equations over the range of the regressor variable(s) used to fit the model. We must be careful if we extrapolate outside of this range.



- **Abuse II**: The disposition of the x values plays an important role in the least-squares fit. While all points have equal weight in determining the height of the line, the slope is more strongly influenced by the remote values of x.



- **Abuse III**: **Outliers** are observations that differ considerably from the rest of the data.
  - They can seriously disturb the least-squares fit.
  - They may be a "bad value" that has resulted from a data recording or some other errors.
  - On the other hand, the data point may not be a bad value and may be a highly useful piece of evidence concerning the process under investigation.



- **Abuse IV**: Though a regression analysis has indicated a strong relationship between two variables, this does not imply that the variables are related in any causal sense.
  - Causal implies necessary correlation.
  - Regression analysis can only address the issues on correlation. It cannot address the issue of necessity.
  - Thus, our expectations of discovering cause-and-effect relationships from regression should be modest.

Example 1.3: Data Illustrating Nonsense Relationships

- **Background**: Table 1.3 presents the number of certified mental defectives in the United Kingdom per 10,000 of estimated population (y), the number of radio receiver licenses issued (x_1), and the first name of the President of the United States (x_2) for the years 1924 – 1937.

Table: Example 1.3

Year	Number of Certified Mental Defectives per 10000 of Estimated Population in the U.K (y)	Number of Radio Receiver Licenses Issued (Millions) in the U.K (x_1)	First Name of President of the U.S. (x_2)
1924	8	1.350	Calvin
1925	8	1.960	Calvin
1926	9	2.270	Calvin
1927	10	2.483	Calvin
1928	11	2.730	Calvin
1929	11	3.091	Calvin
1930	12	3.647	Herbert
1931	16	4.620	Herbert
1932	18	5.497	Herbert
1933	19	6.260	Herbert
1934	20	7.012	Franklin
1935	21	7.618	Franklin
1936	22	8.131	Franklin
1937	23	8.593	Franklin

```
1  data3 <- read.table("example1.3.txt", header=TRUE)
2  res1 <- lm(y~x1, data=data3)
3  summary(res1)
```

```
Call:
lm(formula = y ~ x1, data = data3)

Residuals:
    Min      1Q  Median      3Q     Max
-0.9024 -0.5181 -0.2144  0.4317  1.3014

Coefficients:
            Estimate Std. Error t value Pr(>|t|)
(Intercept)   4.5822     0.4233   10.82 1.51e-07 ***
x1            2.2042     0.0807   27.31 3.58e-12 ***
---
Signif.codes: 0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 0.7262 on 12 degrees of freedom
Multiple R-squared:  0.9842,  Adjusted R-squared:  0.9828
F-statistic:   746 on 1 and 12 DF,  p-value: 3.577e-12
```

- The regression equation relating y and x_1 is $\hat{y} = 4.5822 + 2.2042x_1$.
- The t statistic for testing $H_0 : \beta_1$ for this model is $t_0 = 27.31$ (the p-value is $3.58e^{-12}$).
- The coefficient of determination is $R^2 = 0.9842$. That is 98.42% of the variability in the data is explained by the number of radio receiver licenses issued.

- Remark:
  - This is a **nonsense relationship**, as it is highly unlikely that the number of mental defectives in the population is functionally related to the number of radio receiver licenses issued.
  - Reason: y and x_1 are monotonically related (two sequences of numbers are monotonically related if as one sequence increases the other always either increases or decreases).
  - In this example, y is increasing because diagnostic procedures for mental disorders are becoming more refined over the years represented in the study and x_1 is increasing because of the emergence and low-cost availability of radio technology over the years.

```
1  data3 <- read.table("example1.3.txt", header=TRUE)
2  res2 <- lm(y~x2, data=data3)
3  summary(res2)
```

```
Call:
lm(formula = y ~ x2, data = data3)

Residuals:
    Min      1Q  Median      3Q     Max
-3.7206 -1.4485  0.2574  1.3015  3.2794

Coefficients:
            Estimate Std. Error t value Pr(>|t|)
(Intercept) -26.5882     4.2840  -6.206 4.54e-05 ***
x2            6.0441     0.6202   9.746 4.73e-07 ***
---
Signif.codes: 0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 1.933 on 12 degrees of freedom
Multiple R-squared:  0.8878,  Adjusted R-squared:  0.8785
F-statistic: 94.98 on 1 and 12 DF,  p-value: 4.728e-07
```

- The regression equation relating y and x_2 is $\hat{y} = -26.5882 + 6.0441x_2$.
- The t statistic for testing $H_0 : \beta_1$ for this model is $t_0 = 9.746$ (the p-value is $4.73e^{-7}$).
- The coefficient of determination is $R^2 = 0.8878$. That is 88.78% of the variability in the data is explained by the number of radio receiver licenses issued.
- Remark: This is a nonsense relationship as well.

- **Abuse V**: In some applications of regression the value of the regressor variable x required to predict y is unknown.
  - Consider predicting maximum daily load on an electric power generation system from a regression model relating the load to the maximum daily temperature;
  - To predict tomorrow's maximum load, we must first predict tomorrow's maximum temperature;
  - Consequently, the prediction of maximum load is **conditional** on the temperature forecast;
  - The accuracy of the maximum load forecast depends on the accuracy of the temperature forecast. This must be considered when evaluating model performance.

Regression through the Origin

- **SLR model (through the origin)**: $y = \beta_1 x + \varepsilon$
- **Assumptions**: $\mathrm{E}(\varepsilon_i) = 0$, $\mathrm{Cov}(\varepsilon_i, \varepsilon_j) = \begin{cases} \sigma^2 & i = j \\ 0 & i \neq j \end{cases}$, $i = 1, \ldots, n$.
- Least-squares function: $S(\beta_1) = SS_{Res}(\beta_1) = \sum_{i=1}^{n} \varepsilon_i^2 = \sum_{i=1}^{n} (y_i - \beta_1 x_i)^2$

  Normal equation:

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

- Solution:

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

- The **fitted SLR model through origin** is

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

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

- A $100(1-\alpha)$ percent **confidence interval** (CI) on the slope β_1 is given by

$$\hat{\beta}_1 - t_{\alpha/2,n-1}\sqrt{\frac{MS_{Res}}{\sum_{i=1}^{n} x_i^2}} \leq \beta_1 \leq \hat{\beta}_1 + t_{\alpha/2,n-1}\sqrt{\frac{MS_{Res}}{\sum_{i=1}^{n} x_i^2}}$$

- A $100(1-\alpha)$ percent **confidence interval** (CI) on the **mean response** at the point $x = x_0$ is given by

$$\hat{\mu}_{y|x_0} - t_{\alpha/2,n-1}\sqrt{\frac{x_0^2 MS_{Res}}{\sum_{i=1}^{n} x_i^2}} \leq \mathrm{E}(y|x_0) \leq \hat{\mu}_{y|x_0} + t_{\alpha/2,n-1}\sqrt{\frac{x_0^2 MS_{Res}}{\sum_{i=1}^{n} x_i^2}}$$

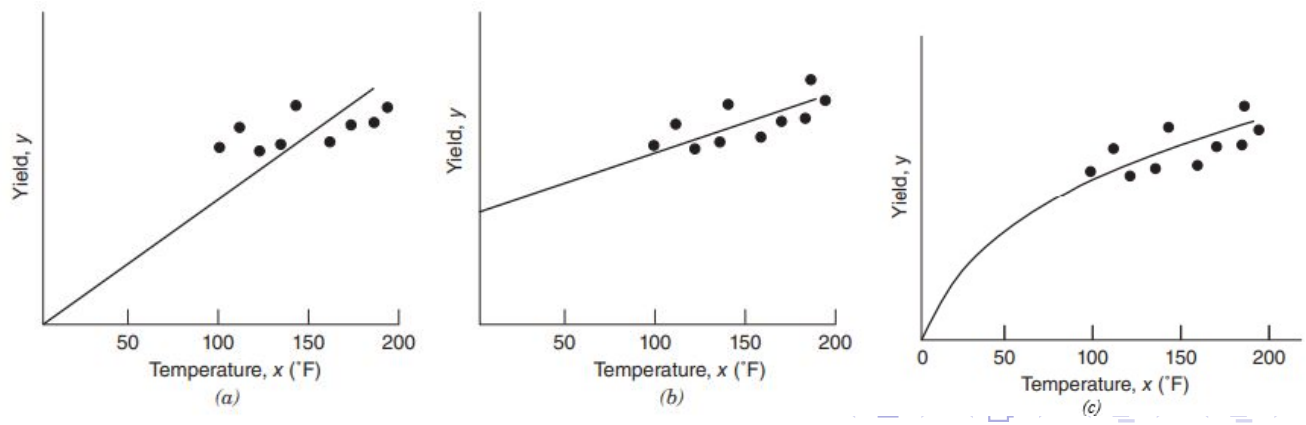
- A $100(1-\alpha)$ percent **prediction interval** on a **future observation** at x_0 is given by

$$\hat{y}_0 - t_{\alpha/2,n-1}\sqrt{MS_{Res}\left(1+\frac{x_0^2}{\sum_{i=1}^{n} x_i^2}\right)} \leq y_0 \leq \hat{y}_0 + t_{\alpha/2,n-1}\sqrt{MS_{Res}\left(1+\frac{x_0^2}{\sum_{i=1}^{n} x_i^2}\right)}$$

Cautions for Regression through the Origin

Caution I: It is relatively easy to misuse the no-intercept model, particularly in situations where the data lie in **a region of x space remote from the origin**.

- Frequently the relationship between y and x is quite different **near the origin** than it is in **the region of x space containing the data**.
  - Although over the range of the regressor variable $100^\circ F \leq x \leq 200^\circ F$, yield and temperature seem to be linearly related, forcing the model to go through the origin provides a visibly poor fit. (figure(a): regression through the origin)
  - A model containing an intercept provides a much better fit in the region of x space where the data were collected. (figure(b): regression with intercept)
  - It would seem that either a quadratic or a more complex nonlinear regression model would be required to adequately express the relationship between y and x over the entire range of x. (figure(c): true relationship)



Caution II: When using regression-through-the-origin model, the **residuals** must be interpreted with care because they usually do not sum to zero.

- From the **normal equations**
  - **(Model with intercept)**

$$n\hat{\beta}_0 + \hat{\beta}_1 \sum_{i=1}^{n} x_i = \sum_{i=1}^{n} y_i$$

$$\hat{\beta}_0 \sum_{i=1}^{n} x_i + \hat{\beta}_1 \sum_{i=1}^{n} x_i^2 = \sum_{i=1}^{n} y_i x_i$$

Constriants on the residuals: $\sum_{i=1}^{n} e_i = 0$, $\sum_{i=1}^{n} x_i e_i = 0$ and $\sum_{i=1}^{n} \hat{y}_i e_i = 0$.

  - **(Model through the origin)**

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

Constraint on the residuals: $\sum_{i=1}^{n} x_i e_i = 0$.

Caution III: Generally R^2 is not a good comparative statistic for the regression model with intercept and through the origin.

- **Partitioning of variability**:
  - **(Model with intercept)**

$$y_i - \bar{y} = (\hat{y}_i - \bar{y}) + (y_i - \hat{y}_i)$$

$$\sum_{i=1}^{n}(y_i - \bar{y})^2 = \sum_{i=1}^{n}(\hat{y}_i - \bar{y})^2 + \sum_{i=1}^{n}(y_i - \hat{y}_i)^2 + 2\sum_{i=1}^{n}(\hat{y}_i - \bar{y})(y_i - \hat{y}_i)$$

$$\sum_{i=1}^{n}(\hat{y}_i - \bar{y})(y_i - \hat{y}_i) = \sum_{i=1}^{n}\hat{\beta}_1(x_i - \bar{x})e_i = \hat{\beta}_1\sum_{i=1}^{n}x_i e_i - \hat{\beta}_1\bar{x}\sum_{i=1}^{n}e_i = 0$$

$$\Rightarrow \sum_{i=1}^{n}(y_i - \bar{y})^2 = \sum_{i=1}^{n}(\hat{y}_i - \bar{y})^2 + \sum_{i=1}^{n}(y_i - \hat{y}_i)^2$$

  - **(Model through the origin)**

$$y_i = (y_i - \hat{y}_i) + \hat{y}_i$$

$$\sum_{i=1}^{n}{y_i}^2 = \sum_{i=1}^{n}(y_i - \hat{y}_i)^2 + \sum_{i=1}^{n}\hat{y}_i^2 + 2\sum_{i=1}^{n}\hat{y}_i(y_i - \hat{y}_i)$$

$$\sum_{i=1}^{n}\hat{y}_i(y_i - \hat{y}_i) = \hat{\beta}_1\sum_{i=1}^{n}x_i e_i = 0$$

$$\Rightarrow \sum_{i=1}^{n}{y_i}^2 = \sum_{i=1}^{n}(y_i - \hat{y}_i)^2 + \sum_{i=1}^{n}\hat{y}_i^2$$

  - Remark: In regression-through-the-origin model, $\sum_{i=1}^{n}(\hat{y}_i - \bar{y})(y_i - \hat{y}_i)$ will generally take a nonzero value and the regression line does not necessarily pass through $(\bar{x}, \bar{y})$.

- **Fundamental analysis-of-variance identity**:
  - **(Model with intercept)**

$$\sum_{i=1}^{n}(y_i - \bar{y})^2 = \sum_{i=1}^{n}(\hat{y}_i - \bar{y})^2 + \sum_{i=1}^{n}(y_i - \hat{y}_i)^2$$

  - **(Model through the origin)**

$$\sum_{i=1}^{n} y_i{}^2 = \sum_{i=1}^{n} \hat{y}_i{}^2 + \sum_{i=1}^{n}(y_i - \hat{y}_i)^2$$

- **Coefficient of determination**:
  - **(Model with intercept)**

$$R^2 = \frac{\sum_{i=1}^{n}(\hat{y}_i - \bar{y})^2}{\sum_{i=1}^{n}(y_i - \bar{y})^2}$$

  - **(Model through the origin)**

$$R_0^2 = \frac{\sum_{i=1}^{n} \hat{y}_i{}^2}{\sum_{i=1}^{n} y_i{}^2}$$

  - Remark: The statistic R_0^2 indicates the **proportion of variability around the origin** (zero) accounted for by regression.

- An alternative way to define R_0^2:

$$R_0^2 = 1 - \frac{\sum_{i=1}^{n}(y_i - \hat{y}_i)^2}{\sum_{i=1}^{n}(y_i - \bar{y})^2}$$

  - In regression-through-the-origin model, $\sum_{i=1}^{n}(y_i - \hat{y}_i)^2$ may exceed $\sum_{i=1}^{n}(y_i - \bar{y})^2$.This can occur when the data form a curvilinear pattern or a linear pattern with an intercept away from the origin. Hence, R_0^2 may turn out to be **negative**. Consequently, the coefficient of determination has no clear meaning for regression through the origin.

- We prefer to use MS_{Res} as a basis of comparison between intercept and no-intercept regression models.

Example 1.4: Shelf-stocking Data

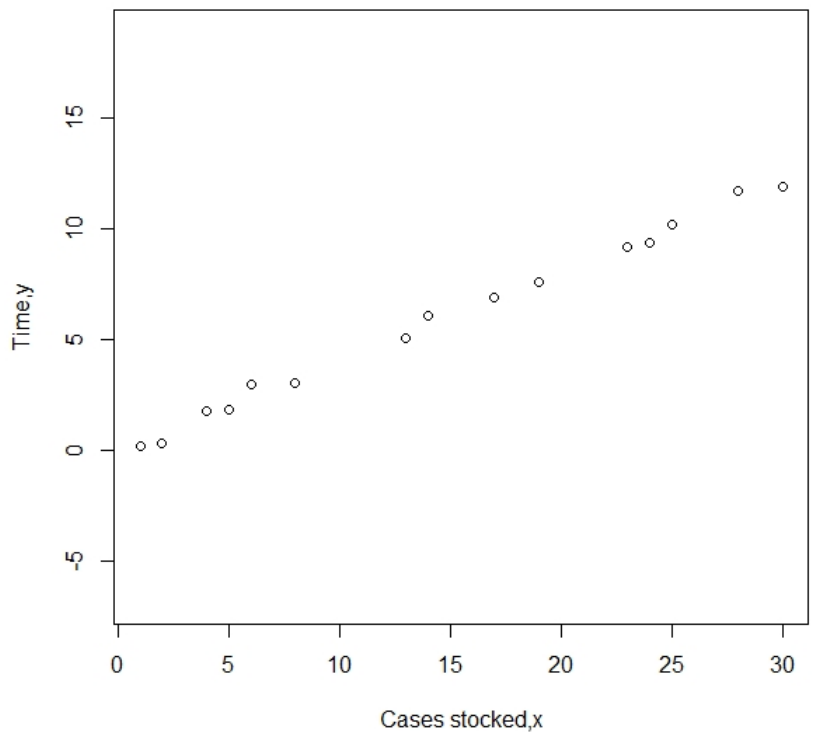
Table: Example 1.4

Times,y(minutes)	Cases Stocked,x
10.15	25
2.96	6
3.00	8
6.88	17
0.28	2
5.06	13
9.14	23
11.86	30
11.69	28
6.04	14
7.57	19
1.74	4
9.38	24
0.16	1
1.84	5

- **Background**: A merchandiser stocked a grocery store shelf with a soft drink product. It is suspected that the time required is related to the number of cases of product stocked. Fifteen observations on the time required and the cases stocked are shown in table example 1.4.
- y: the time required

  x: the number of cases of product stocked
- **Question arising**: Obtain and plot the fitted least-squares equation.

```
data4 <- read.table("example1.4.txt", header=TRUE)
# Scatterplot of the data
plot(data4$x, data4$y, xlab="Cases stocked,x", ylab="Time,y")
```

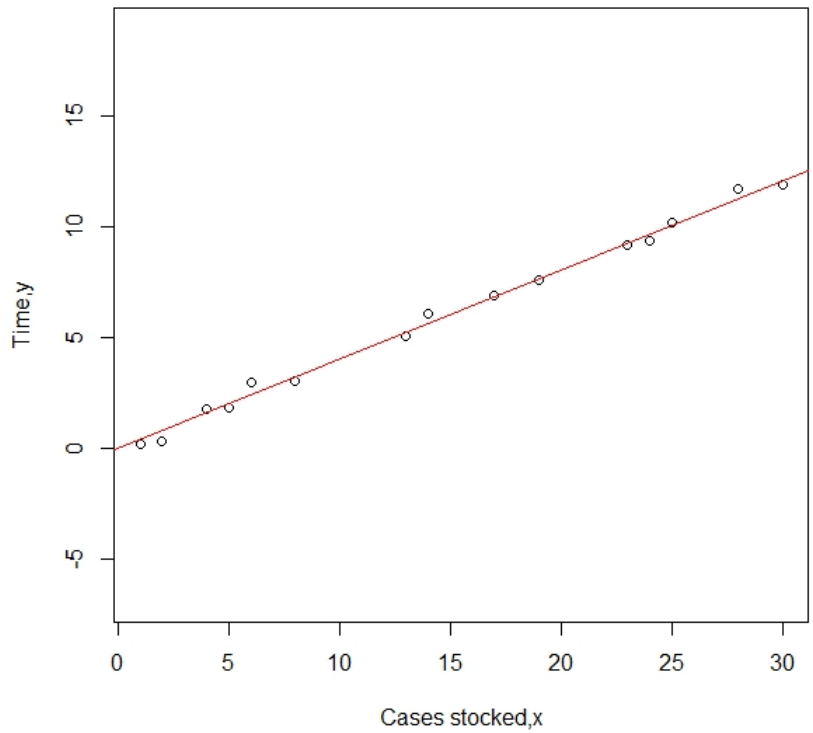


```
1  res <- lm(y~0+x, data=data4)
2  summary(res)
```

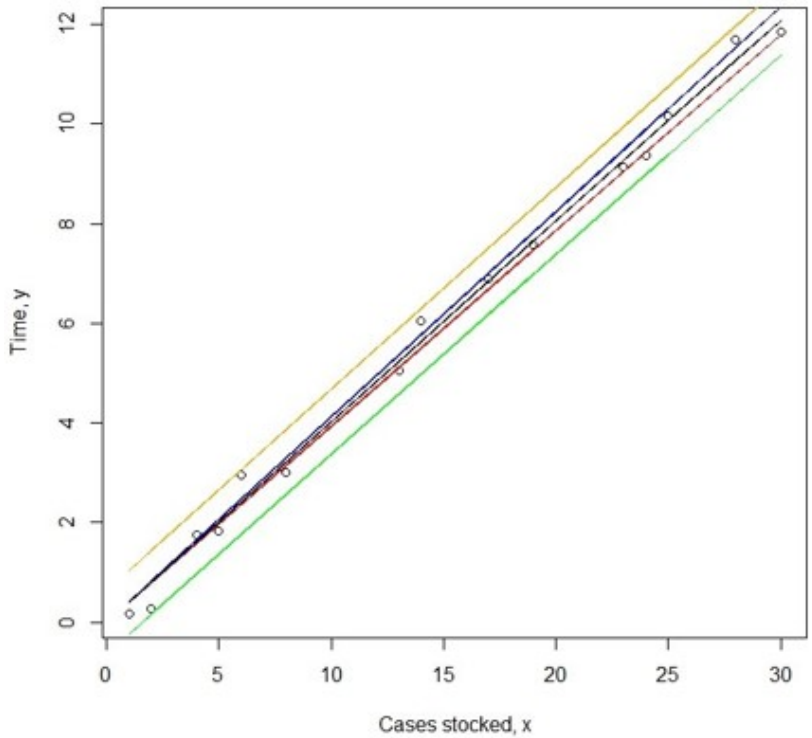
```
Call:
lm(formula = y ~ 0 + x, data = data4)

Residuals:
    Min      1Q  Median      3Q     Max
-0.5252 -0.2198 -0.1202  0.1070  0.5443

Coefficients:
  Estimate Std. Error t value Pr(>|t|)
x 0.402619   0.004418   91.13   <2e-16 ***
---
Signif.codes: 0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 0.2988 on 14 degrees of freedom
Multiple R-squared:  0.9983,  Adjusted R-squared:  0.9982
F-statistic:  8305 on 1 and 14 DF,  p-value: < 2.2e-16
```

- The least-squares fit is $\hat{y} = 0.4026x$

```
1    abline(res, col="red")
```



Confidence & Prediction Intervals through the Origin



Yellow line: Upper 95% prediction limits

Blue line: Upper 95% confidence limits

Red line: Lower 95% confidence limits

Green line: Lower 95% prediction limits

Estimation by Maximum Likelihood

- The method of LS can be used to estiamte parameteres in a LR model **regardless of the form of the distributions of the errors**. If the form of the distribution of the errors is known, the method of maximum Likelihood can be used with likelihood function:

$$L(y_i, x_i, \beta_0, \beta_1, \sigma^2) = \prod_{i=1}^{n} (2\pi\sigma^2)^{-1/2} exp\left[-\frac{1}{2\sigma^2}(y_i - \beta_0 - \beta_1 x_i)^2\right]$$

$$= (2\pi\sigma^2)^{-n/2} exp\left[-\frac{1}{2\sigma^2}\sum_{i=1}^{n}(y_i - \beta_0 - \beta_1 x_i)^2\right]$$

- The **maximum-likelihood estimators** (**MLE**)

$$\tilde{\beta}_0 = \bar{y} - \tilde{\beta}_1 \bar{x}, \quad \tilde{\beta}_1 = \frac{\sum_{i=1}^{n} y_i (x_i - \bar{x})}{\sum_{i=1}^{n} (x_i - \bar{x})^2};$$

$$\tilde{\sigma}^2 = \frac{\sum_{i=1}^{n} (y_i - \tilde{\beta}_0 - \tilde{\beta}_1 x_i)^2}{n}$$

LSE vs. MLE

- For simple linear regression model with **normal errors**, the MLEs are identical to the LSEs of the coefficients.
- MLE have **better statistical properties** than LSE. Maximum likelihood estimation is asymptotically optimal when estimating unknown parameters of a model. When the sample size n is large, it is guaranteed to perform better than any other estimation methods.
- However, MLE requires **more stringent statistical assumptions** than LSE.