Chapter 2

Simple Linear Regression

Summary

Simple linear regression is used to study possible relationship between a response variable and a regressor (or explanatory) variable.

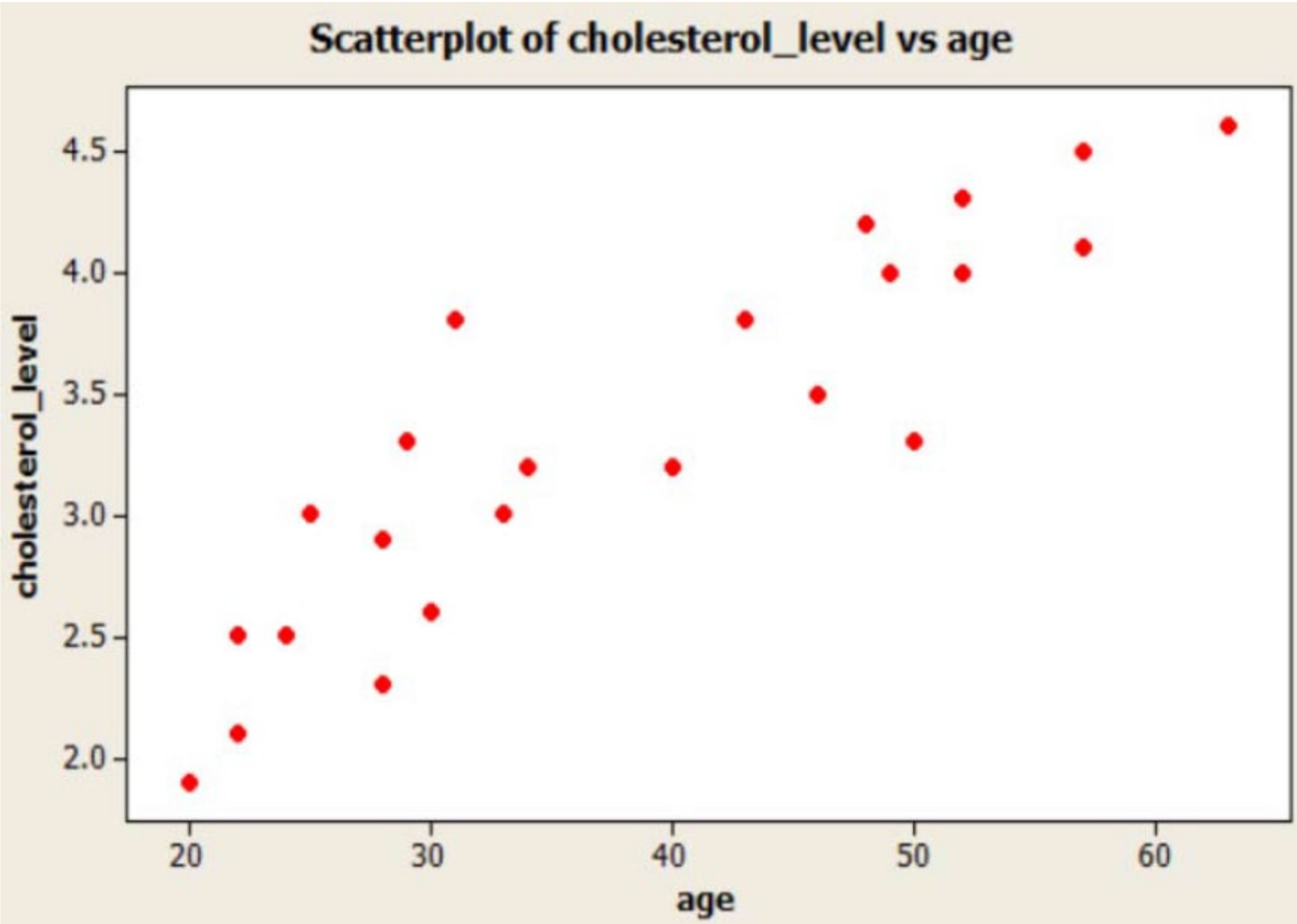
Examples

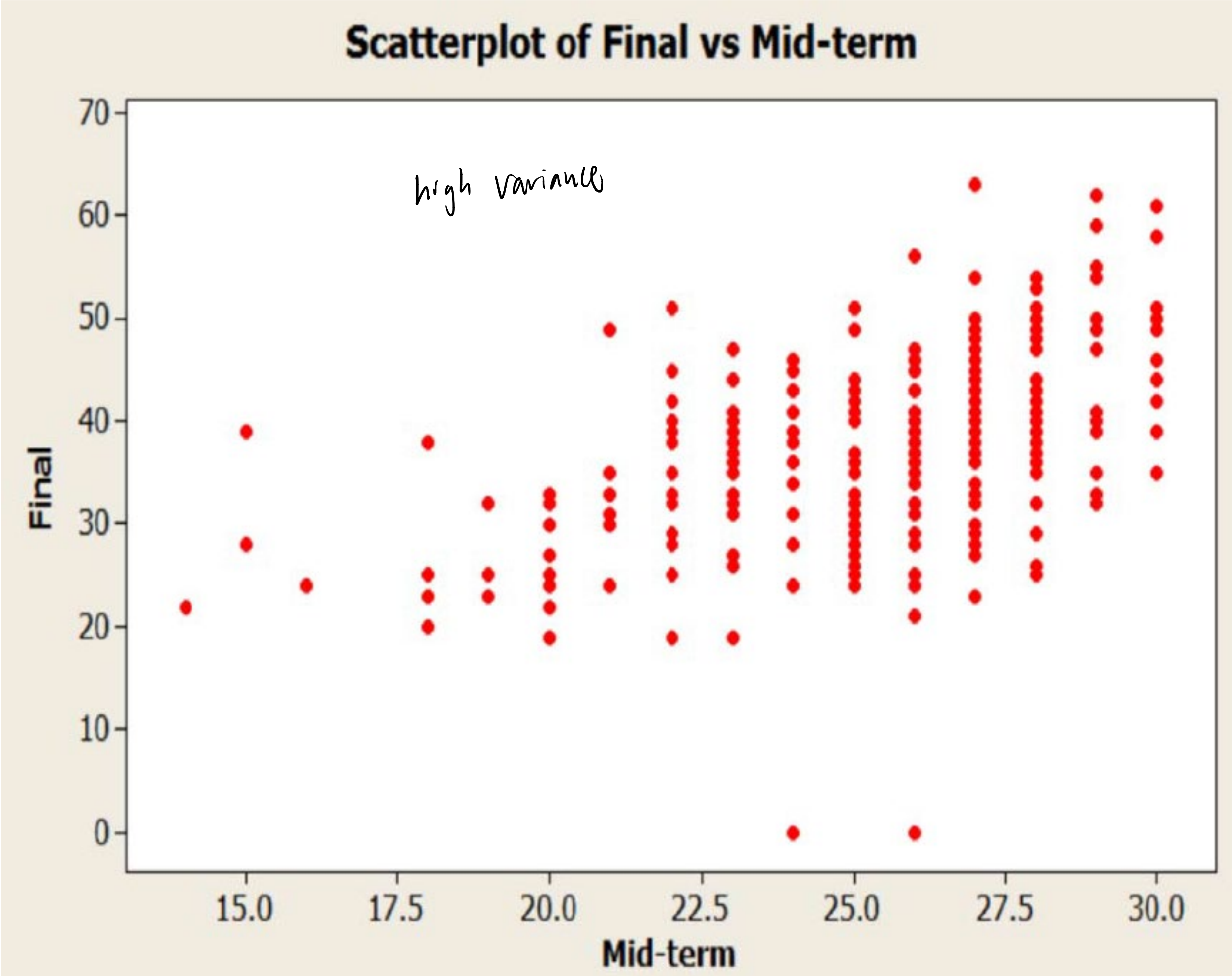
response variable

explanatory variable

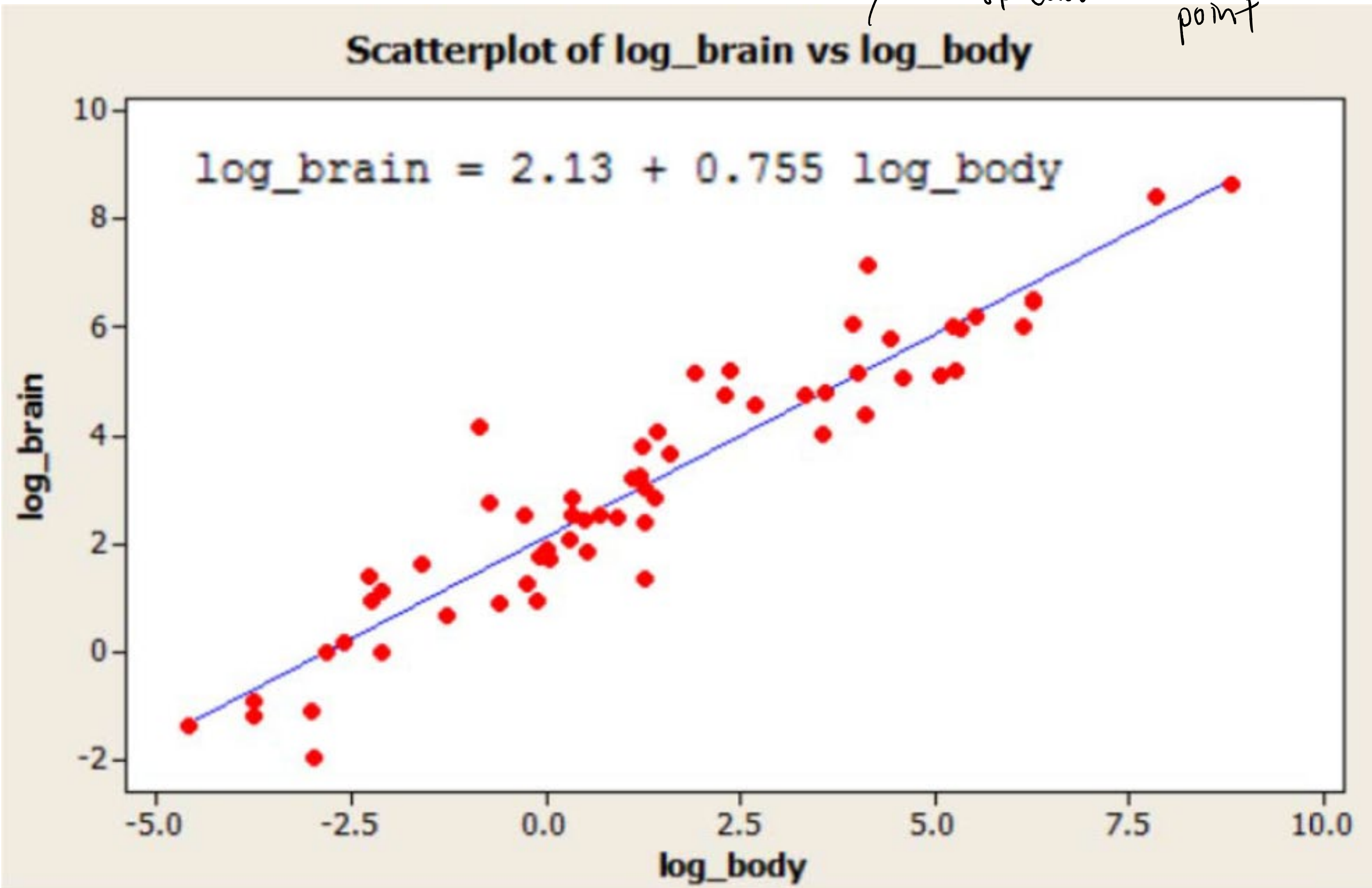
- Cholesterol level versus age
- Final exam score versus mid-term exam score
- Log(brain weight) versus Log(body weight)
- Lifetime versus length of lifeline

The response variable is usually plotted on the y-axis and the regressor variable on the x-axis.


Scatterplot of cholesterol_level vs age
cholesterol_level
4.5
4.0
3.5
3.0
2.5
2.0
20
30
40
50
60
age


Scatterplot of Final vs Mid-term
high variance
Final
70
60
50
40
30
20
10
0
15.0
17.5
20.0
22.5
25.0
27.5
30.0
Mid-term

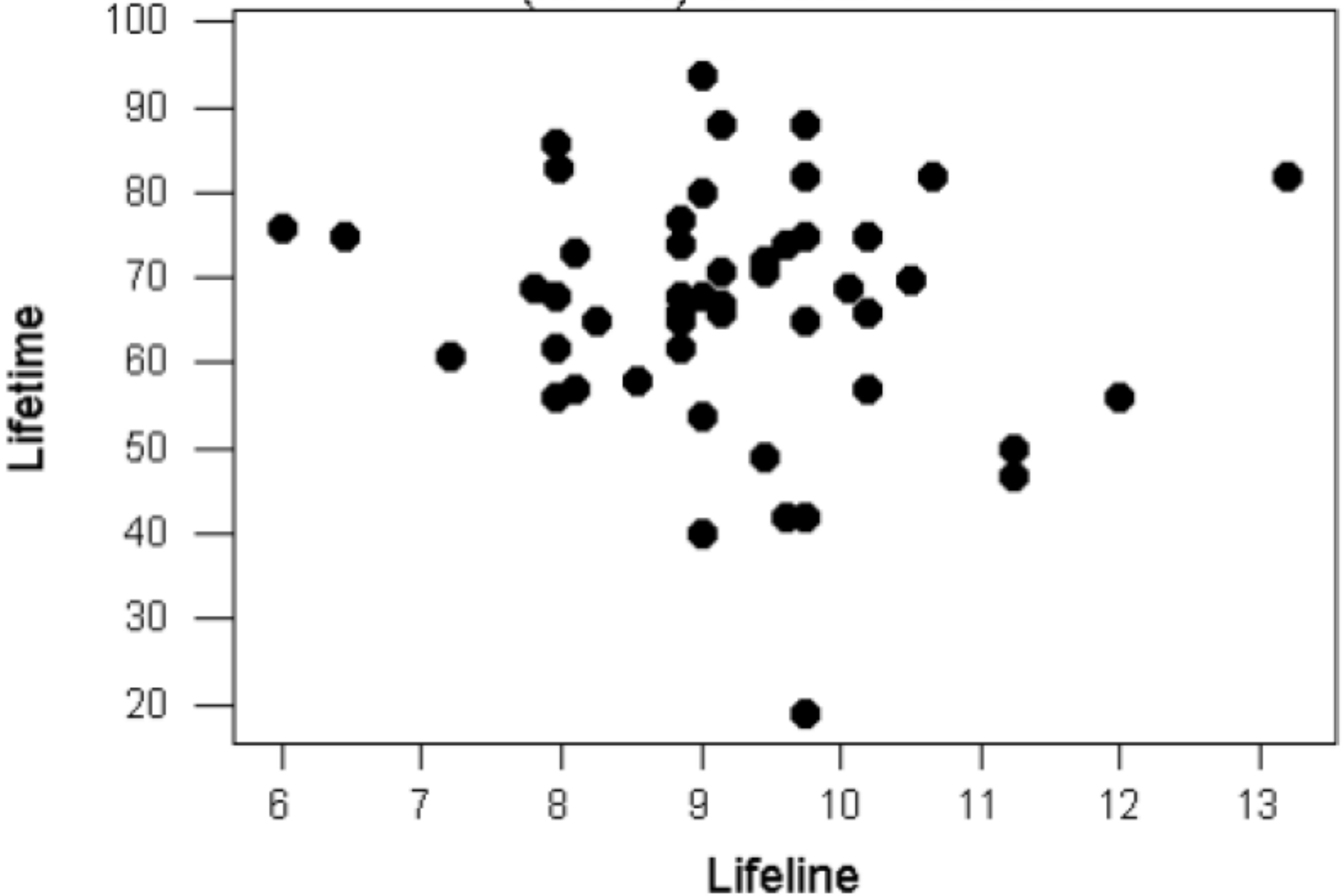
log used to
spread out data
point



Is there any relationship between lifeline and lifetime?

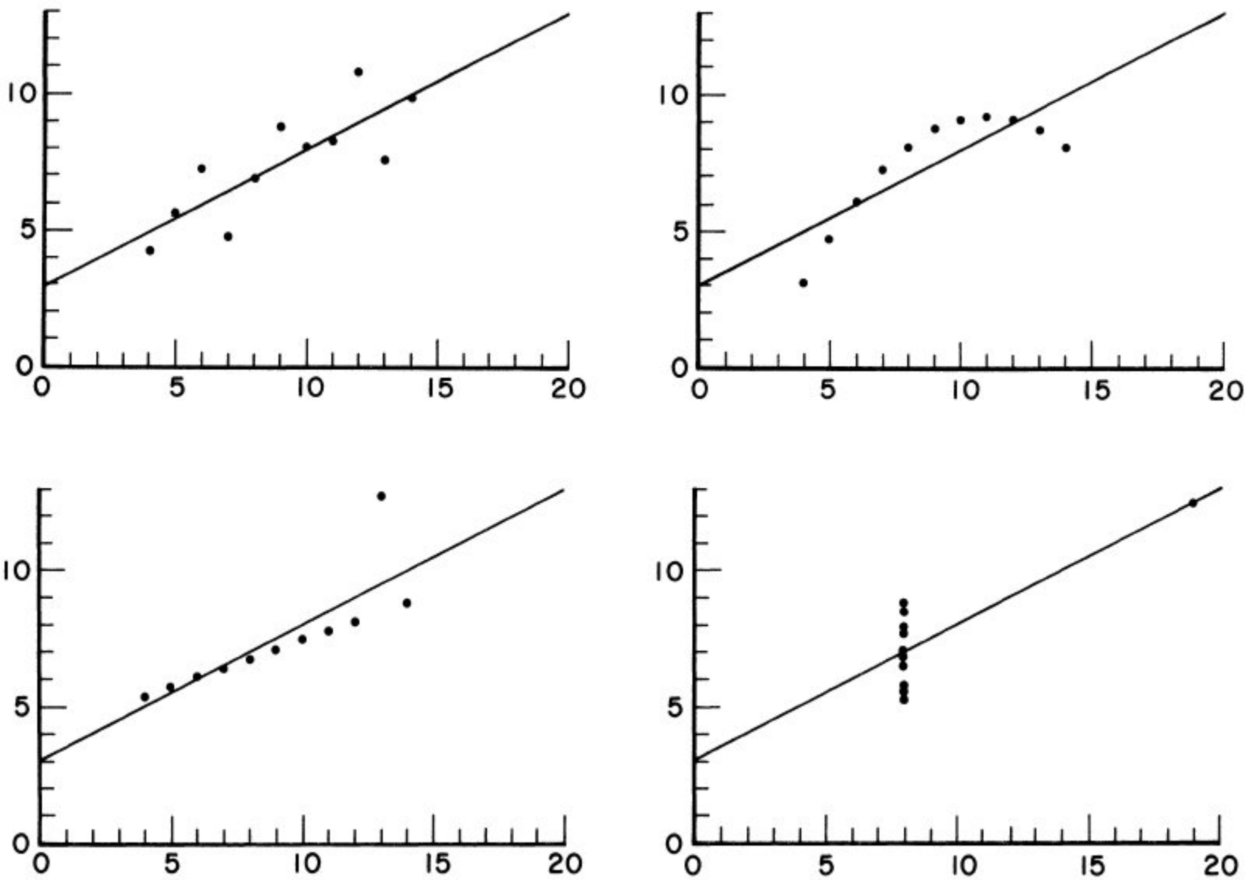
↓	C1	C2
	age	lifeline
1	19	9.75
2	40	9.00
3	42	9.60
4	42	9.75
5	47	11.25
6	49	9.45
7	50	11.25
8	54	9.00
9	56	7.95
10	56	12.00
11	57	8.10
12	57	10.20
13	58	8.55
14	61	7.20
15	62	7.95
16	62	8.85
17	65	8.25
18	65	8.85
19	65	9.75
20	66	8.85

A Study to Investigate Possible Relationship Between the Lifetime (in Years) and Lifeline (in cm) of Individuals



Importance of plotting

best fitted line
↳ same straight line
for all 4 different
data sets



Graphs in Statistical Analysis* F. J. ANSCOMBE

The American Statistician, Vol. 27, No. 1 (Feb., 1973), pp. 17–21

Simple linear regression model

The simple linear regression model for a response variable y and a regressor variable x based on observations (x_1, y_1), (x_2, y_2), ..., (x_n, y_n) can be written as

y_i is random variable

eplison used to denote the error

$$\underbrace{y_i = \beta_0 + \beta_1 x_i}_{} + \boxed{\epsilon_i,}\ i = 1, 2, ..., n,$$

data point flactuates above and below best fitted line

where ϵ_i is a random variable such that $E(\epsilon_i) = 0$, $Var(\epsilon_i) = \sigma^2$, $i = 1, 2, ..., n$ and ϵ_i's are independent.

average, expected error $= 0$

true for $\forall i$ $\Rightarrow$ Uniform variation (constant for $\forall x$)

Example - Rocket propellant data

A rocket motor is manufactured by bonding an igniter propellant and a sustainer propellant together inside a metal casing.

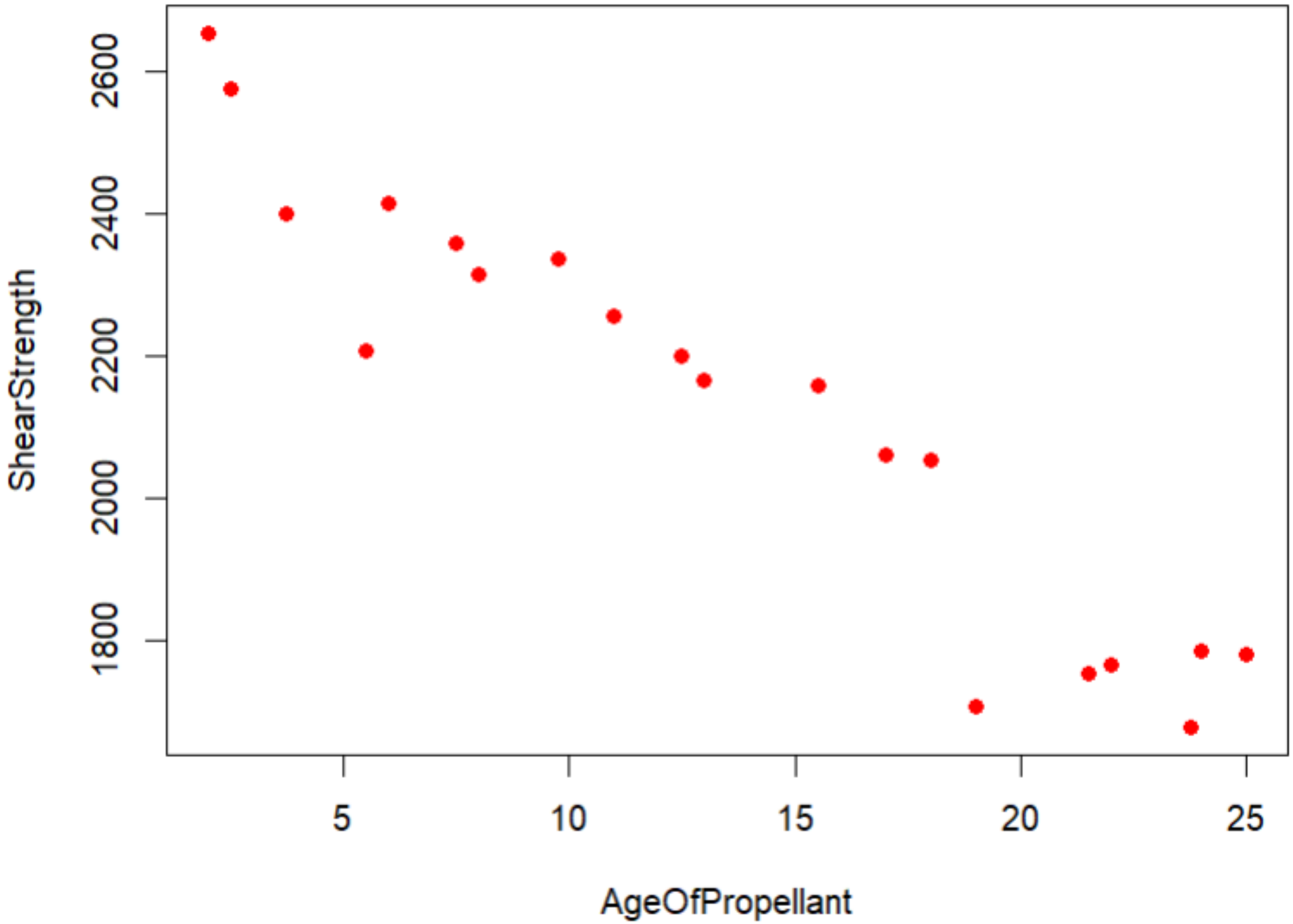
The shear strength of the bond between the two types of propellant is an important quality characteristic.

It is suspected that shear strength (y) is related to the age (x) in weeks of the batch of sustainer propellant.

Twenty observations were collected.

Rocket_Prop.csv - Excel

Observation, i	Shear Strength, yi (psi) (y)	Age of Propellant, xi (weeks) (x)
1	2158.7	15.5
2	1678.15	23.75
3	2316	8
4	2061.3	17
5	2207.5	5.5
6	1708.3	19
7	1784.7	24
8	2575	2.5
9	2357.9	7.5
10	2256.7	11
11	2165.2	13
12	2399.55	3.75
13	1779.8	25
14	2336.75	9.75
15	1765.3	22
16	2053.5	18
17	2414.4	6
18	2200.5	12.5
19	2654.2	2
20	1753.7	21.5



ShearStrength
2600
2400
2200
2000
1800
5
10
15
20
25
AgeOfPropellant

R can be downloaded for free:

https://cran.r-project.org/bin/windows/base/

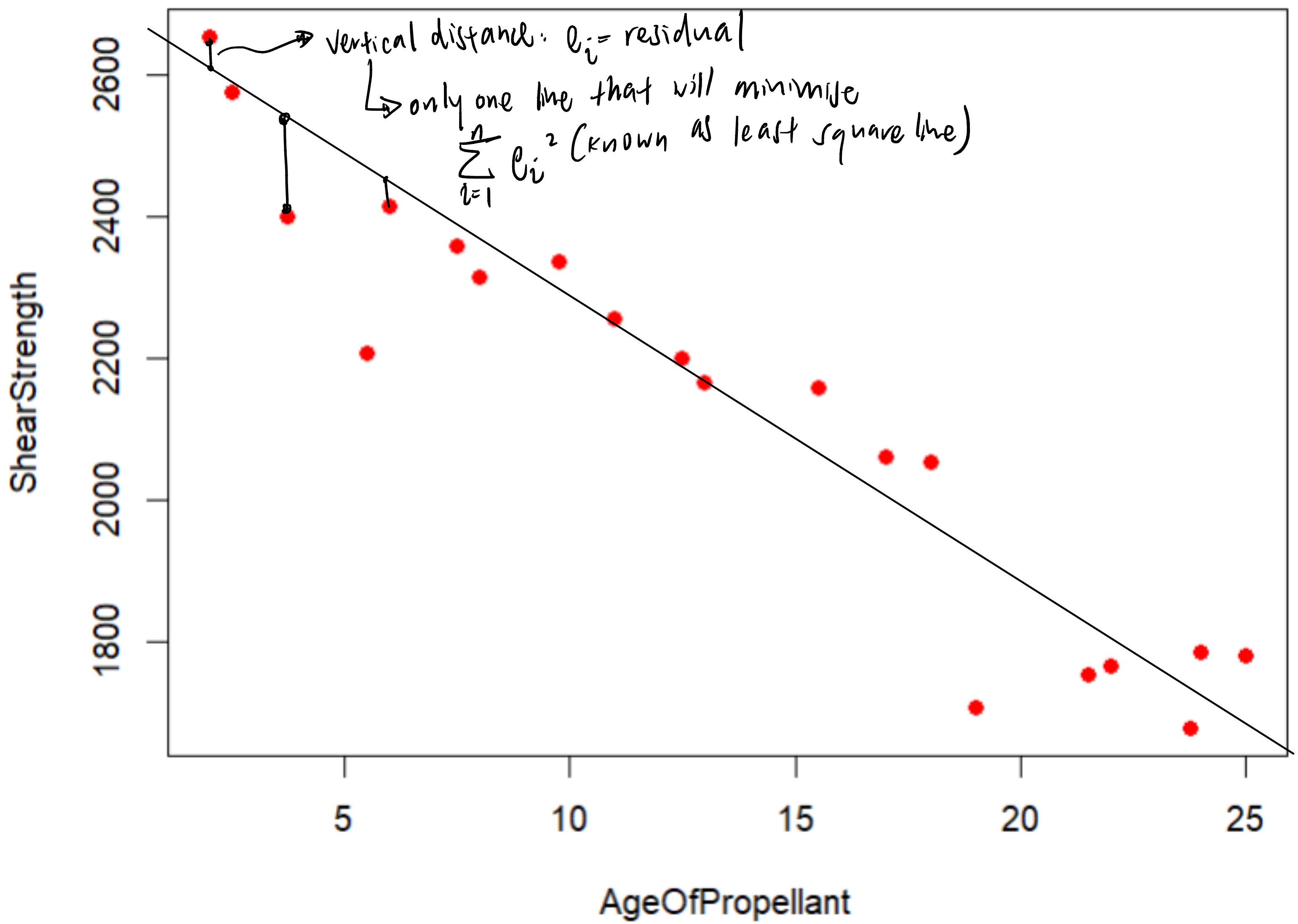
https://cran.r-project.org/bin/macosx/

RStudio can be downloaded for free:

https://posit.co/download/rstudio-desktop/

Example – Plotting of rocket propellant data using R

```r
# ch2_rocket_prop.R
#
library(MASS)
rm(list = ls())
rocket.data <- read.table("D:\\nus_teaching\\st3131\\data\\Rocket_Prop.csv",
                          header = T, sep=",")
rocket.data

#choose simpler names for the two variables
names(rocket.data) <- c("Obs", "ShearStrength", "AgeOfPropellant")
rocket.data

#attach() function is used to access variables present in the dataframe
attach(rocket.data)
y <- rocket.data[,2]
x <- rocket.data[,3]
nobs <- length(x)

#plot the data
plot(x,y,pch=16,col="red")
plot(AgeOfPropellant, ShearStrength,pch=16,col="red")
```



vertical distance: e_i = residual
only one line that will minimise $\sum_{i=1}^{n} e_i^2$ (known as least square line)
2600
2400
2200
2000
1800
ShearStrength
5
10
15
20
25
AgeOfPropellant

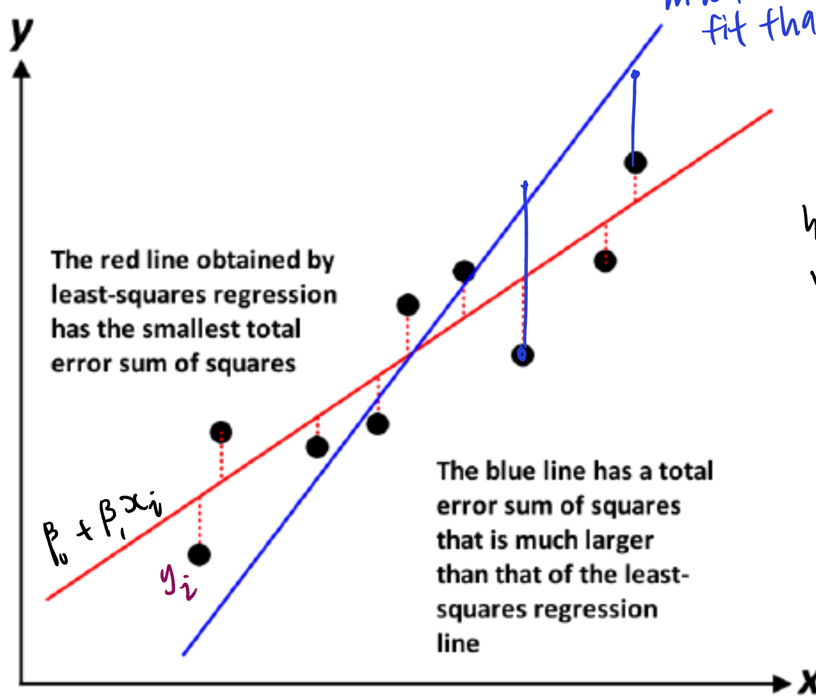
Least-squares function and least-squares regression

The least-squares function is the error sum of squares

$$S = \sum_{i=1}^{n} \epsilon_i^2 = \sum_{i=1}^{n} \left(y_i - \beta_0 - \beta_1 x_i \right)^2 .$$

Least-squares regression means fitting a model that minimizes the error sum of squares. This method is appealing because the fitted model is the closest to the data in terms of error sum of squares.

Refer to figure for a comparison between the least-squares line and a non least-squares line. The least-squares line (red) is much closer to the data than the non least-squares line (blue).



much worse fit than the red line (eg. blue line)

hence only ONE LINE will be the least square line

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

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

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

least-squares normal equations

minimise function

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

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

least-squares estimators $\hat{\beta}_0$ and $\hat{\beta}_1$

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

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

sample mean

fitted simple linear regression model

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

straight line without error

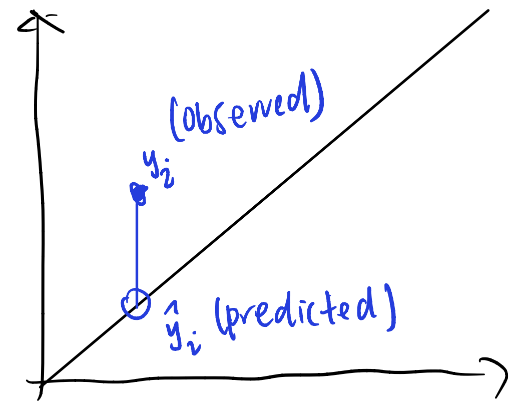
Simpler notations

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

$$S_{xy} = \sum_{i=1}^{n} y_i x_i - \frac{\left(\sum_{i=1}^{n} y_i\right)\left(\sum_{i=1}^{n} x_i\right)}{n} = \sum_{i=1}^{n} y_i (x_i - \bar{x})$$

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

$$\hat{\beta}_1 = \frac{S_{xy}}{S_{xx}}$$



*i*th residual

$$e_i = y_i - \hat{y}_i = y_i - \left(\hat{\beta}_0 + \hat{\beta}_1 x_i\right), \quad i = 1, 2, \ldots, n$$

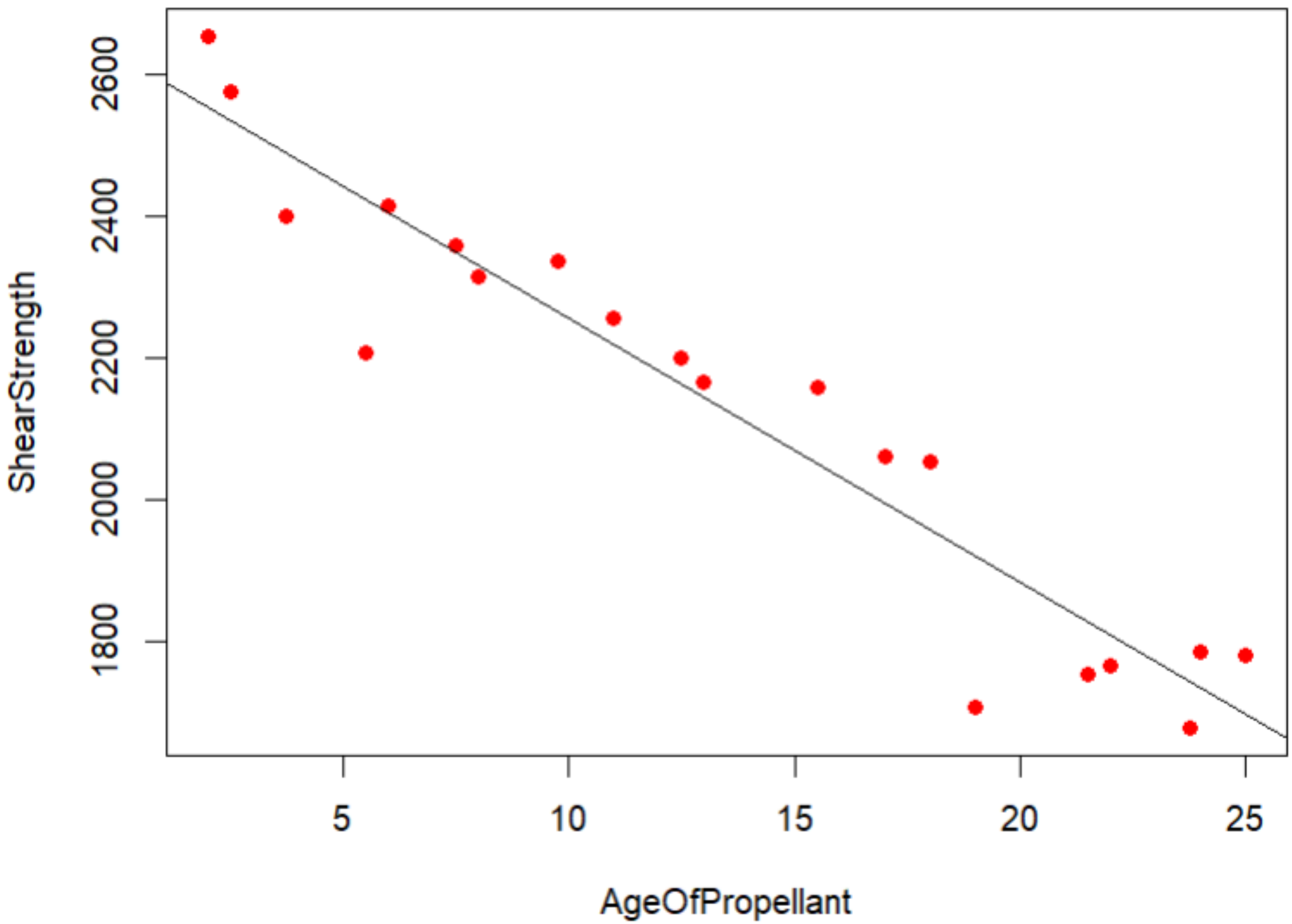
observed

predicted

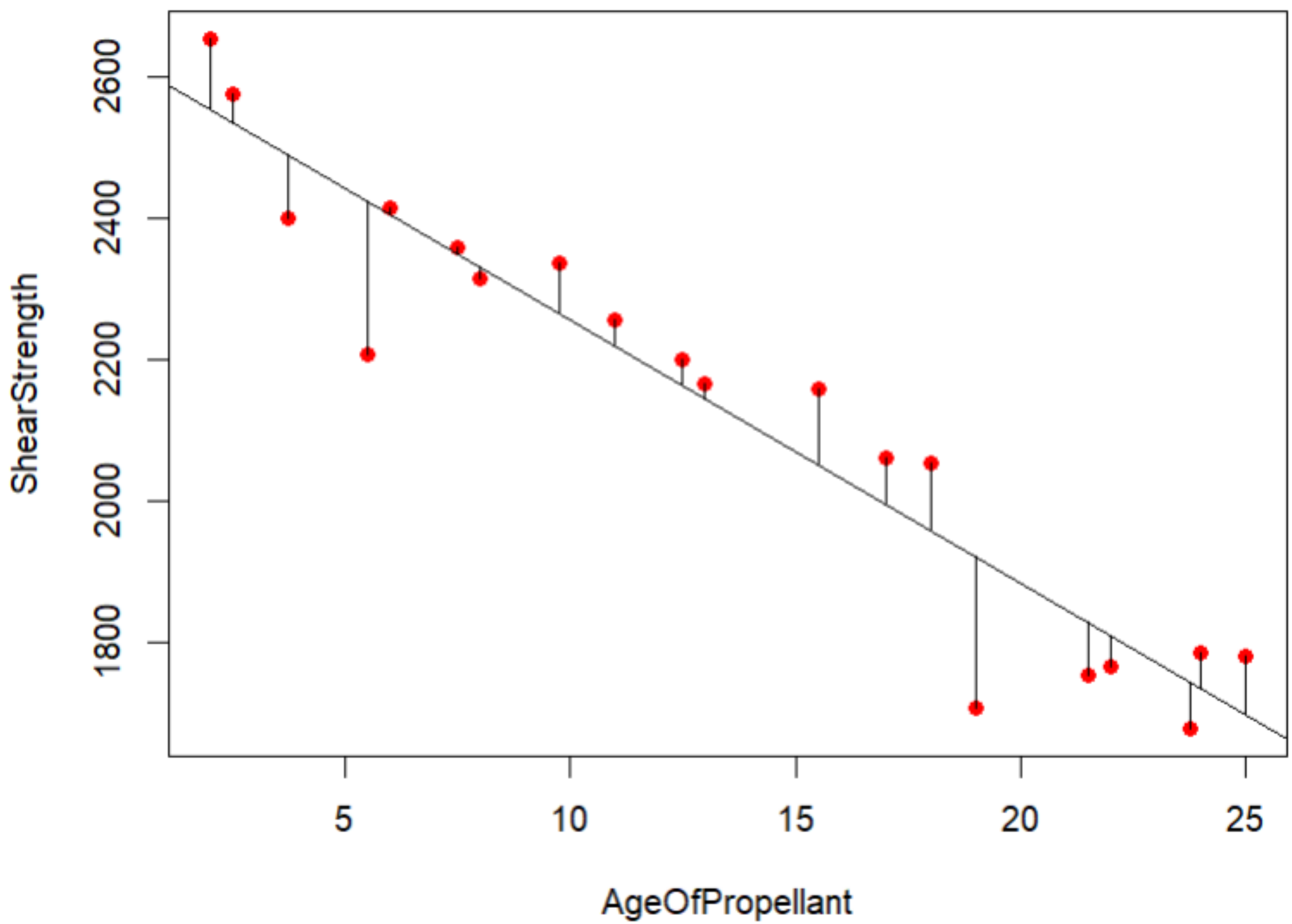
Example – Analysis of rocket propellant data using R

```
#fit linear regression line
fitted.model <- lm(y~x)
summary(fitted.model)

#obtain beta0 and beta1
COEF <- coef(fitted.model)
names(COEF)
names(COEF) <- NULL
beta0 <- COEF[1]
beta1 <- COEF[2]
beta0
beta1

#plot least-squares line
abline(beta0,beta1)

#calculate fitted values of x
yhat <- predict(fitted.model) #calculate fitted values of x

#plot residuals
for (i in 1:nobs) lines(c(x[i],x[i]),c(y[i],yhat[i]))
```

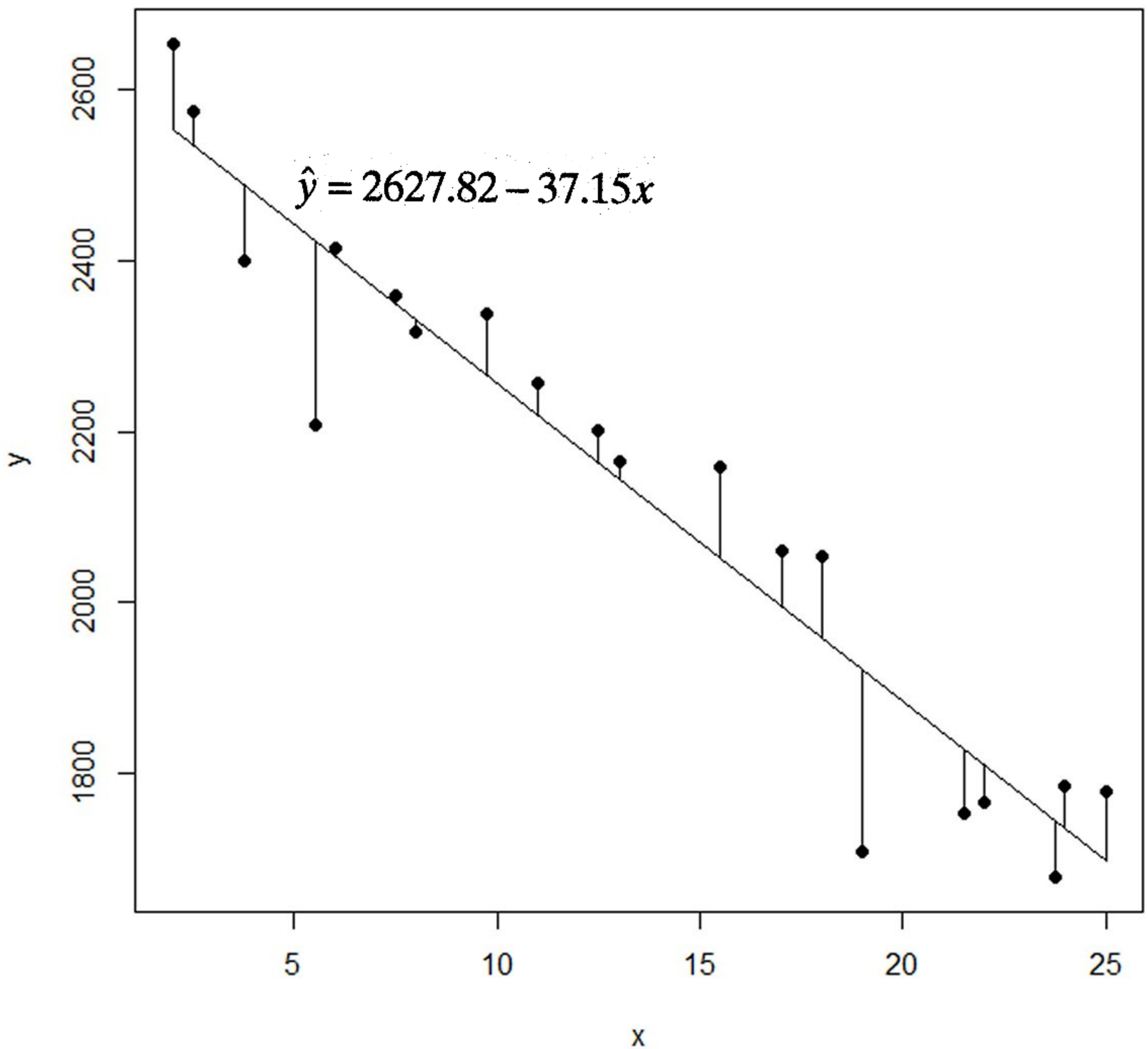
```
> #fit linear regression line
> fitted.model <- lm(y~x)
> summary(fitted.model)

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
```



ShearStrength
2600
2400
2200
2000
1800
5
10
15
20
25
AgeOfPropellant



ShearStrength
2600
2400
2200
2000
1800
5
10
15
20
25
AgeOfPropellant



$\hat{y} = 2627.82 - 37.15x$
y
x
1800
2000
2200
2400
2600
5
10
15
20
25

The End