1. Simple Linear Regression

- **Regression** used to study the relationship between quantitative variables (e.g. height and weight)
- Collect data on two or more variables
- Two variables – **simple linear regression** (Chapter 1)
- More than two variables – **multiple regression** (Chapter 2)
- Response variable (Y) = variable to be predicted
- Predictor variable (X) = variable used to predict Y
- (x_i, y_i) = observed values of X and Y for ith case.
- Number of cases = n

Height (X)	**Weight (Y)**
x_1	y_1
x_2	y_2
.	.
.	.
.	.
x_n	y_n

Outline of simple linear regression

- Assume there is a linear relationship between X and Y: $Y = \beta_0 + \beta_1 X$
- β_0 is the intercept (value of Y when X = 0) and β_1 is the slope (change in Y for a unit change in X)
- Estimate β_0 and β_1 from the data
- Use model to predict Y for given X

Methods of linear regression

- Draw **scatter plot** of points (x_i, y_i); does relationship look approximately linear?
- If so, find the **line of closest fit** to the points; estimate β_0 and β_1 from the data
- How well does the model fit? Any outliers?
- Would other models give better fit?

  $Y = \beta_0 + \beta_1 X + \beta_2 X^2$ (Quadratic in X)?

  $Y = \beta_0 + \beta_1 X + \beta_2 Z$, where Z = Age (Multiple regression)?

Forbes Data Set

- Atmospheric pressure (measured by mercury barometer) determines altitude
- Low pressure corresponds to high altitude & high pressure corresponds to low altitude

- Use boiling point of water as alternative to measuring atmospheric pressure.
- As atmospheric pressure ↑, boiling point of water ↑ (e.g. pressure cooker)
- As atmospheric pressure ↓, boiling point of water ↓

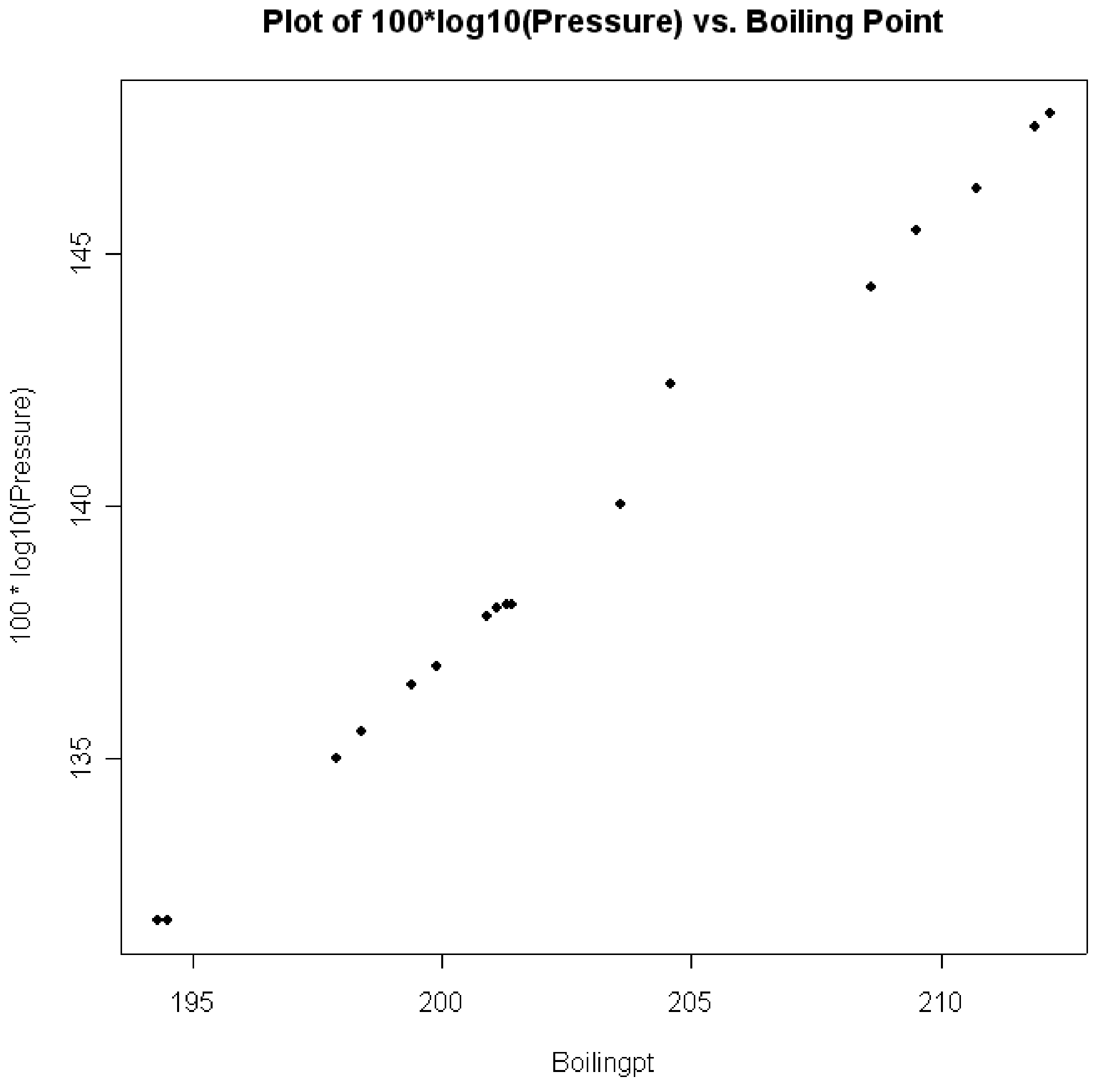
- Use boiling point of water to predict pressure
- Forbes(1857) measured boiling point of water and pressure at 17 locations
- Theory suggests a model of the form

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

  where Y = 100 log(Pressure),
  and Pressure is measured in inches of mercury (Hg)
  and X = Boiling point of water in °F

Case	Boiling Pt (^{o}F)	Pressure (in Hg)	Log(Pressure)	100xLog(Pressure)
1	194.5	20.79	1.3179	131.79
2	194.3	20.79	1.3179	131.79
3	197.9	22.40	1.3502	135.02
4	198.4	22.67	1.3555	135.55
5	199.4	23.15	1.3646	136.46
6	199.9	23.35	1.3683	136.83
7	200.9	23.89	1.3782	137.82
8	201.1	23.99	1.3800	138.00
9	201.4	24.02	1.3806	138.06
10	201.3	24.01	1.3804	138.04
11	203.6	25.14	1.4004	140.04
12	204.6	26.57	1.4244	142.44
13	209.5	28.49	1.4547	145.47
14	208.6	27.76	1.4434	144.34
15	210.7	29.04	1.4630	146.30
16	211.9	29.88	1.4754	147.54
17	212.2	30.06	1.4780	147.80

Scatter plot for Forbes data

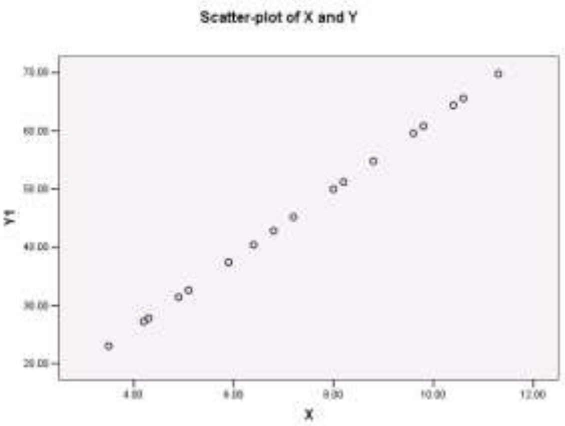


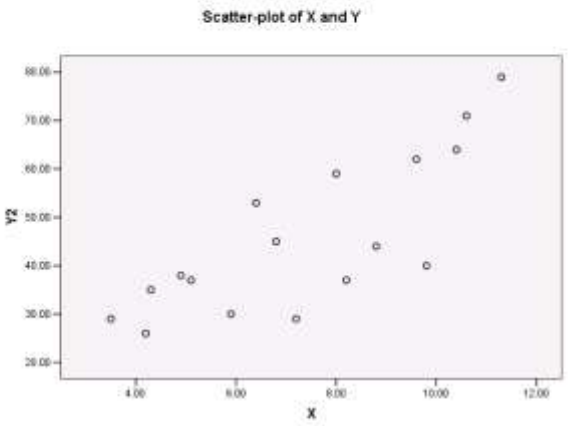
- The points generally fall on a straight line
- There is one “outlier”

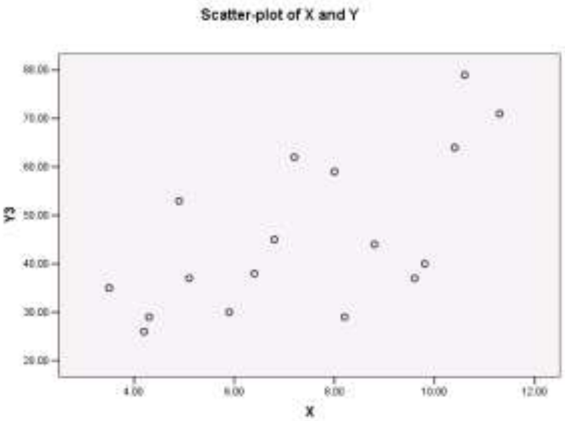
🎥 Video 1.1

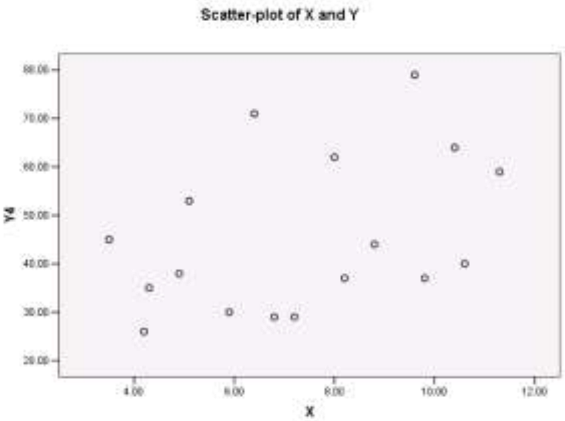
Examples of Scatter plots:

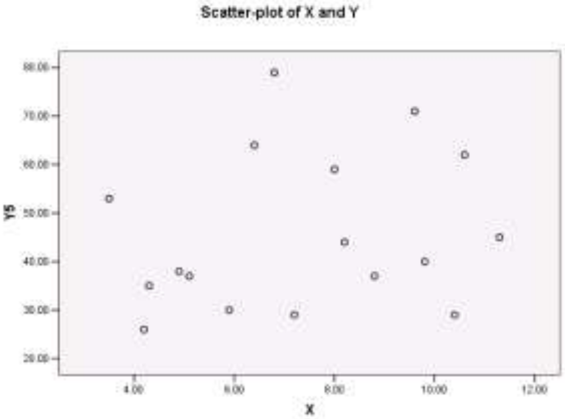
X	Y1	Y2	Y3	Y4	Y5	Y6	Y7	Y8
3.5	23.0	29	35	45	53	40	-23.0	-29
4.2	27.2	26	26	26	26	29	-27.2	-26
4.3	27.8	35	29	35	35	59	-27.8	-35
4.9	31.4	38	53	38	38	71	-31.4	-38
5.1	32.6	37	37	53	37	35	-32.6	-37
5.9	37.4	30	30	30	30	37	-37.4	-30
6.4	40.4	53	38	71	64	45	-40.4	-53
6.8	42.8	45	45	29	79	29	-42.8	-45
7.2	45.2	29	62	29	29	62	-45.2	-29
8.0	50.0	59	59	62	59	64	-50.0	-59
8.2	51.2	37	29	37	44	53	-51.2	-37
8.8	54.8	44	44	44	37	38	-54.8	-44
9.6	59.6	62	37	79	71	37	-59.6	-62
9.8	60.8	40	40	37	40	79	-60.8	-40
10.4	64.4	64	64	64	29	30	-64.4	-64
10.6	65.6	71	79	40	62	26	-65.6	-71
11.3	69.8	79	71	59	45	44	-69.8	-79

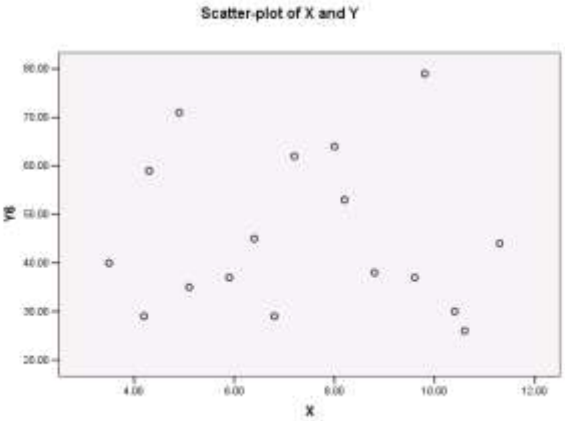


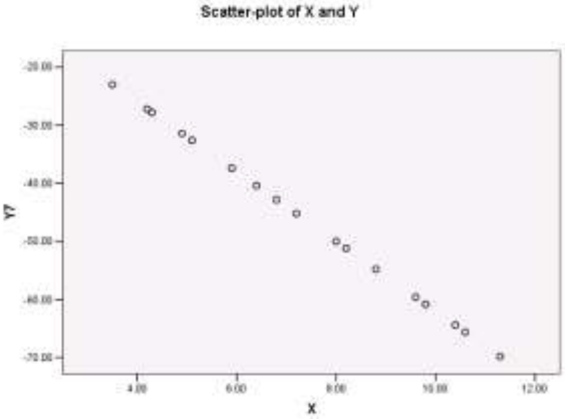


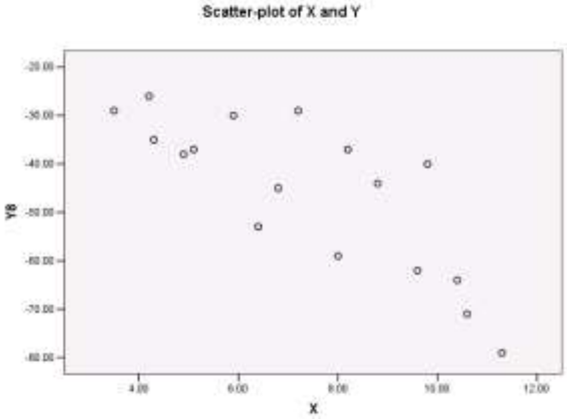

Scatter-plot of X and Y


Scatter-plot of X and Y


Scatter-plot of X and Y


Scatter-plot of X and Y


Scatter-plot of X and Y


Scatter-plot of X and Y

Video 1.2

Assumptions about errors

Real data will almost never fall exactly on a straight line. For most (or all) points there will be an error. These errors could be

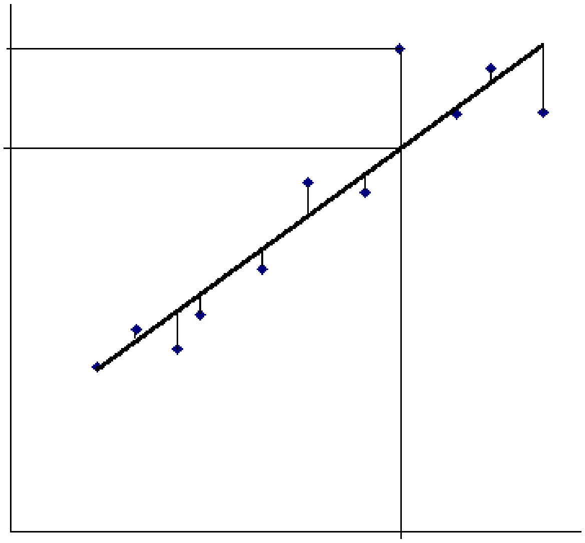
- Measurement errors – impossible to measure continuous variables completely accurately.
- The effect of variables not included in the model.
- Natural variability.

We incorporate these errors into the **simple linear regression model** as follows:

$y_i = \beta_0 + \beta_1 x_i + e_i$, where e_i = **error** for the ith case

and

$y_i = \beta_0 + \beta_1 x_i$ is the **true regression line** of Y on X



We make **assumptions** about errors e_i:

- needed to prove the optimality of the estimates for β_0 and β_1
- needed to find confidence intervals for β_0 and β_1
- tests for these assumptions in Chapter 3/4 (**regression diagnostics**)

$e_i \sim \text{NID}(0, \sigma^2)$

e_i are **Normally** distributed with mean 0

e_i have **common variance** σ^2

e_i are **independent** variables

Assumptions about e_i can be expressed in terms of covariance

$E(e_i) = 0$, $\text{var}(e_i) = \sigma^2$, $\text{cov}(e_i,e_j) = 0$, for $i \neq j$

With the normality assumption, this implies that the e_i's are independent

When applying a regression model, these assumptions have to be verified.

🎥 Video 1.3

Covariance and independent variables

A and B are **independent events** if P(A|B) =P(A)
i.e., knowing that B has occurred doesn't affect the probability of A
Equivalently, A and B are independent events if P(A and B) = P(A)P(B)

X and Y are **independent discrete variables** if
$P(X = x \text{ and } Y = y) = P(X = x)P(Y = y)$

X and Y are **independent continuous variables** if
joint pdf of X and Y = h(x,y) = $f_X(x)\, g_Y(y)$ - the product of individual pdfs

Covariance of X and Y = cov(X,Y) = $E(X - \mu_1)(Y - \mu_2)$, where μ_1= E(X) and μ_2= E(Y)

If X and Y are **independent**, then **cov(X,Y) = 0**:

$$\text{cov}(X,Y) = E(X - \mu_1)(Y - \mu_2) = \int (x - \mu_1)\, f_X(x)dx \bullet \int (y - \mu_2)\, g_y(y)dy = 0$$

If (X,Y) have a bivariate **normal** distribution,
then **cov(X,Y) = 0** ⇒ X,Y **independent**

In general, cov(X,Y) measures the **association** between X and Y,
i.e. the extent to which X and Y vary together

Large X occurs with large Y and small X occurs with small Y ⇔ **positive** association
Large X occurs with small Y and small X occurs with large Y ⇔ **negative** association

Sample covariance $= \dfrac{1}{n-1}\sum (x_i - \bar{x})(y_i - \bar{y})$ estimates cov(X,Y)

If $(x_i - \bar{x}) > 0$ and $(y_i - \bar{y}) > 0$ then $(x_i - \bar{x})(y_i - \bar{y}) > 0$
If $(x_i - \bar{x}) < 0$ and $(y_i - \bar{y}) < 0$ then $(x_i - \bar{x})(y_i - \bar{y}) > 0$
Positive association between X and Y ⇔ sample covariance > 0

If $(x_i - \bar{x}) > 0$ and $(y_i - \bar{y}) < 0$ then $(x_i - \bar{x})(y_i - \bar{y}) < 0$
If $(x_i - \bar{x}) < 0$ and $(y_i - \bar{y}) > 0$ then $(x_i - \bar{x})(y_i - \bar{y}) < 0$
Negative association between X and Y ⇔ sample covariance < 0

Sign of sample covariance indicates the direction of association: positive or negative

Properties of sample correlation coefficient r_{XY}

$$r_{XY} = \frac{SXY}{\sqrt{(SXX)(SYY)}} = \frac{SXY / (n-1)}{\sqrt{(SXX / (n-1))(SYY / (n-1))}}$$

Where
$$SXY = \sum(x_i - \bar{x})(y_i - \bar{y}) = \sum x_i y_i - n\bar{x}\,\bar{y}$$
$$SXX = \sum(x_i - \bar{x})^2 = \sum x_i^2 - n(\bar{x})^2$$
$$SYY = \sum(y_i - \bar{y})^2 = \sum y_i^2 - n(\bar{y})^2$$

Correlation coefficient r_{XY} is the sample covariance, scaled to lie in [-1, 1]

$-1 <= r_{XY} <= 1$

$r_{XY} > 0 \Leftrightarrow$ positive association; $r_{XY} < 0 \Leftrightarrow$ negative association

$r_{XY} = 1 \Leftrightarrow$ all points lie on line with positive slope

$r_{XY} = -1 \Leftrightarrow$ all points lie on line with negative slope

The closer r_{XY} is to +1, the closer the points are to a line with positive slope

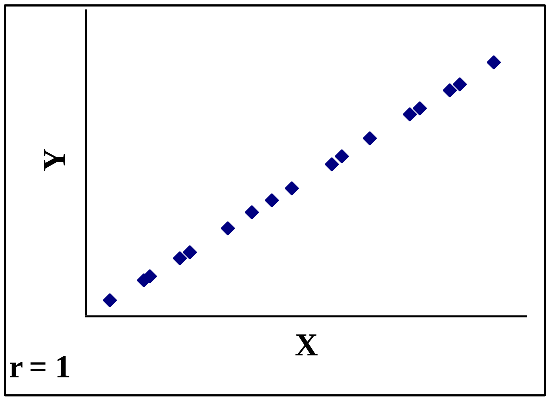
The closer r_{XY} is to -1, the closer the points are to a line with negative slope

In simple regression, r (or its square r^2) is used to measure how well the linear model fits the data

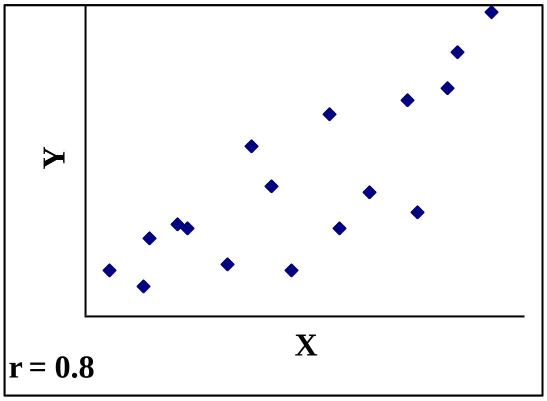
In multiple regression, the multiple correlation coefficient (R^2) is used to measure how well the linear model fits the data

For the Forbes data,
$\bar{x} = 202.95$ $\bar{y} = 139.61$
SXX = 530.78 SXY = 475.30 SYY = 427.76
(You should verify these results using a calculator/spreadsheet/R as an exercise)

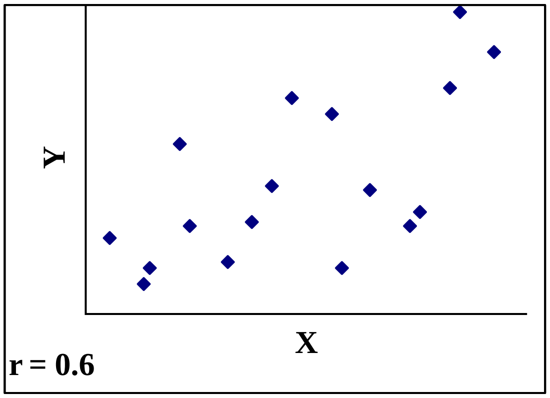
$$r_{XY} = \frac{475.30}{\sqrt{(530.78)(427.76)}} = 0.9975$$



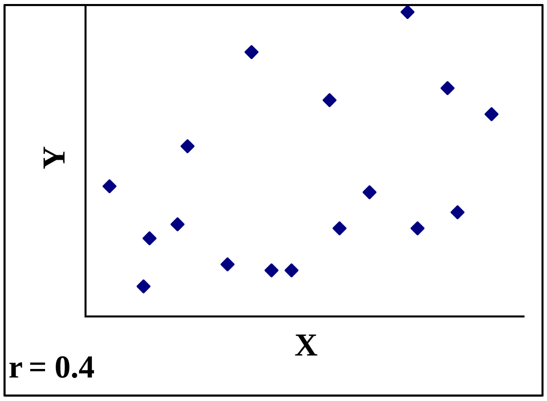
r = 1



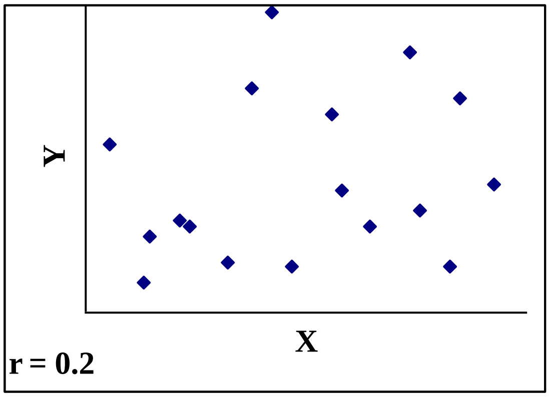
r = 0.8



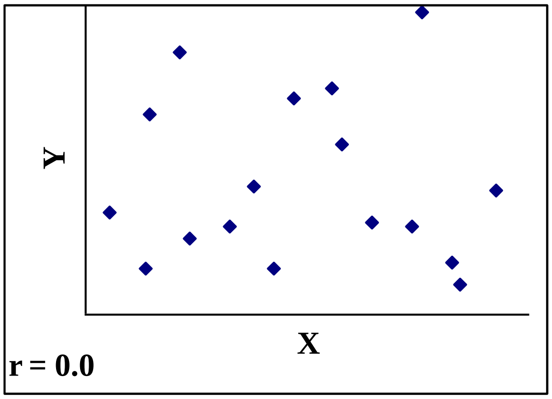
r = 0.6



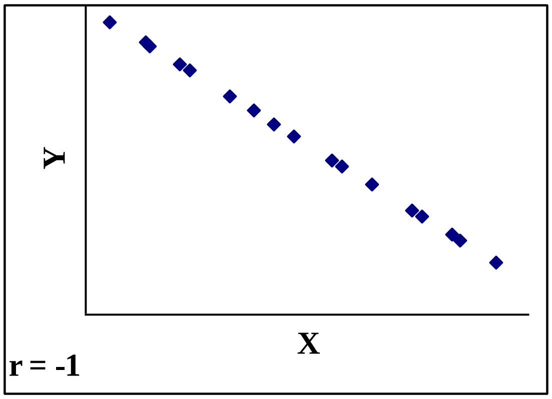
r = 0.4



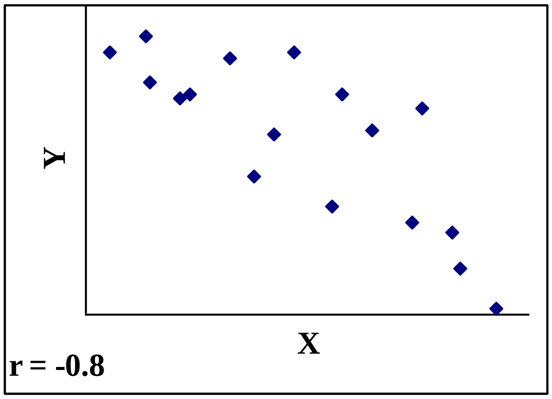
r = 0.2



r = 0.0



r = -1



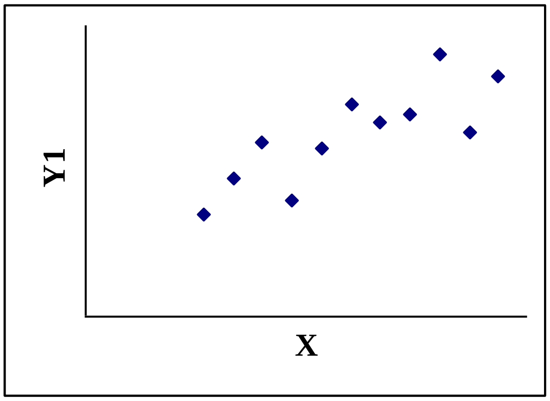
r = -0.8

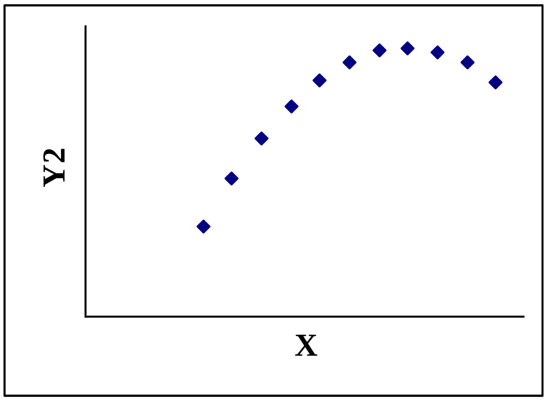
🎥 Video 1.4

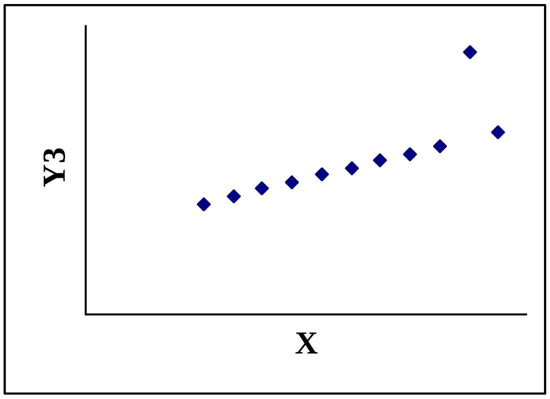
Linearity cannot be deduced from the correlation coefficient. The correlation coefficient should never be interpreted in isolation – but in conjunction with the scatter-plot: Consider the correlation between the Y's and the X for the following data.

X	**Y1**	**Y2**	**Y3**
10	8.04	9.14	7.46
8	6.95	8.14	6.77
13	7.58	8.74	12.74
9	8.81	8.77	7.11
11	8.33	9.26	7.81
14	9.96	8.10	8.84
6	7.24	6.13	6.08
4	4.26	3.10	5.39
12	10.84	9.13	8.15
7	4.82	7.26	6.42
5	5.68	4.74	5.73
r =	0.82	0.82	0.82

All give the same r – this may lead us to believe that each of these Y's has the same relationship with X.
But consider the scatter-plots:

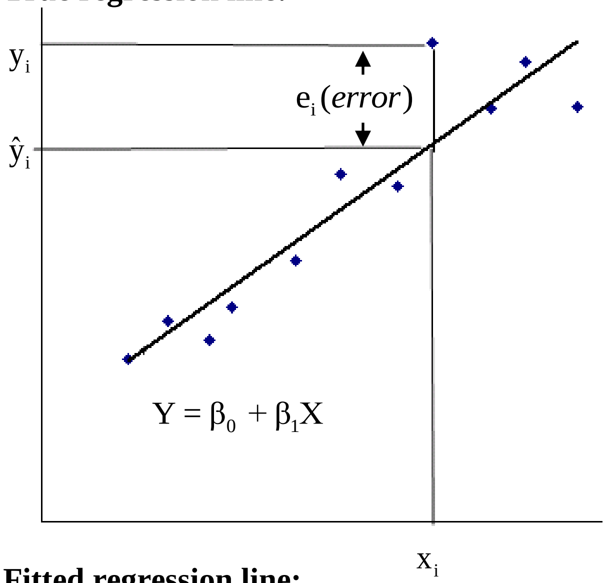




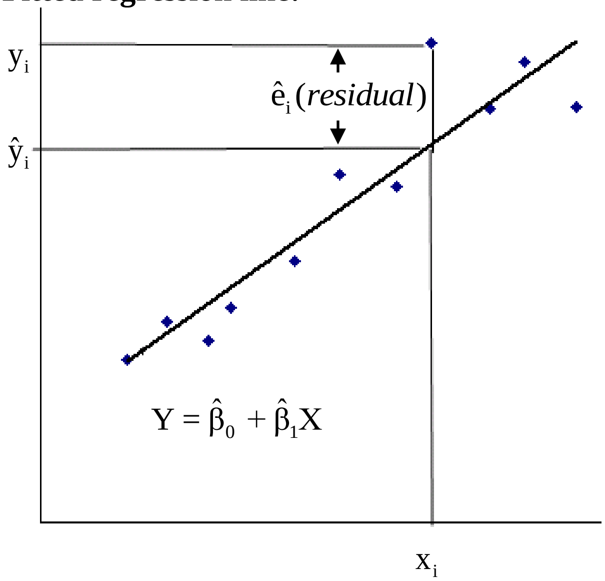


Least Squares Criterion:
In a scatter plot there may be many potential lines that could be fitted to the data. The method for choosing the best line is the Least Squares Criterion.

True regression line:



Fitted regression line:



Chose $\hat{\beta}_0$ and $\hat{\beta}_1$ to minimize the sum of the squared residuals ($\sum_{i=1}^{n} \hat{e}_i^2$).

RSS (Residual Sum of Squares) is the (minimized) value of $\sum_{i=1}^{n} \hat{e}_i^2$.

Through partial differentiation, the following estimators can be derived:

$$\hat{\beta}_1 = \frac{SXY}{SXX} \qquad \hat{\beta}_0 = \bar{y} - \hat{\beta}_1 \bar{x}$$

Video 1.5

Derivation of regression coefficient estimators (for ST2053 and ST6018 only)

$RSS = \sum_{i=1}^{n} \hat{e}_i^2 = \sum (y_i - (\hat{\beta}_0 + \hat{\beta}_1 x_i))^2$ Note: the index on the summation sign is dropped for simplicity.

To minimize RSS, we need to differentiate with respect to 2 unknowns – we need the partial derivatives.
We set these partial derivatives equal to zero and solve the (simultaneous) equations.
(Remember that y_i and x_i are known (the data) and that $\hat{\beta}_0$ and $\hat{\beta}_1$ are the unknowns.)

$$\frac{\delta RSS}{\delta \hat{\beta}_0} = 2\sum (y_i - \hat{\beta}_0 - \hat{\beta}_1 x_i)(-1) = 0$$

$$\sum (y_i - \hat{\beta}_0 - \hat{\beta}_1 x_i) = 0$$

$$\sum y_i = \sum \hat{\beta}_0 + \sum \hat{\beta}_1 x_i$$

$$\sum y_i = n\hat{\beta}_0 + \hat{\beta}_1 \sum x_i$$

This is called the 1^{st} normal equation.

$$\frac{\delta RSS}{\delta \hat{\beta}_1} = 2\sum (y_i - \hat{\beta}_0 - \hat{\beta}_1 x_i)(-x_i) = 0$$

$$\sum (x_i y_i - \hat{\beta}_0 x_i - \hat{\beta}_1 x_i^2) = 0$$

$$\sum x_i y_i = \sum \hat{\beta}_0 x_i + \sum \hat{\beta}_1 x_i^2$$

$$\sum x_i y_i = \hat{\beta}_0 \sum x_i + \hat{\beta}_1 \sum x_i^2$$

This is called the 2^{nd} normal equation.

To solve these simultaneous equations:
Multiply each term in the 1st normal equation by $-\sum x_i$ and every term in the 2nd normal equation by n and add:

$$\begin{array}{l} -\sum x_i \sum y_i = -n\hat{\beta}_0 \sum x_i - \hat{\beta}_1 (\sum x_i)^2 \\ \underline{n\sum x_i y_i = \qquad n\hat{\beta}_0 \sum x_i + n\hat{\beta}_1 \sum x_i^2} \end{array}$$

$$n\sum x_i y_i - \sum x_i \sum y_i = n\hat{\beta}_1 \sum x_i^2 - \hat{\beta}_1 (\sum x_i)^2$$ (This eliminated $\hat{\beta}_0$)

$$n\sum x_i y_i - \sum x_i \sum y_i = \hat{\beta}_1 (n\sum x_i^2 - (\sum x_i)^2)$$

Thus $\hat{\beta}_1 = \dfrac{n\sum x_i y_i - \sum x_i \sum y_i}{n\sum x_i^2 - (\sum x_i)^2}$ Remember $\bar{x} = \dfrac{\sum x_i}{n}$, similarly for $\bar{y}$

$\hat{\beta}_1 = \dfrac{n\sum x_i y_i - n\bar{x}n\bar{y}}{n\sum x_i^2 - (n\bar{x})^2}$ Can cancel n:

$$\hat{\beta}_1 = \frac{\sum x_i y_i - n\bar{x}\;\bar{y}}{\sum x_i^2 - n(\bar{x})^2} = \frac{\mathbf{SXY}}{\mathbf{SXX}}$$

Consider the 1st normal equation:

$$\sum y_i = n\hat{\beta}_0 + \hat{\beta}_1 \sum x_i$$

$n\hat{\beta}_0 = \sum y_i - \hat{\beta}_1 \sum x_i$ Divide each term by n:

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

Note: The proof that the solution to the normal equations provides a minimum rather than a maximum is omitted.

For the Forbes data,

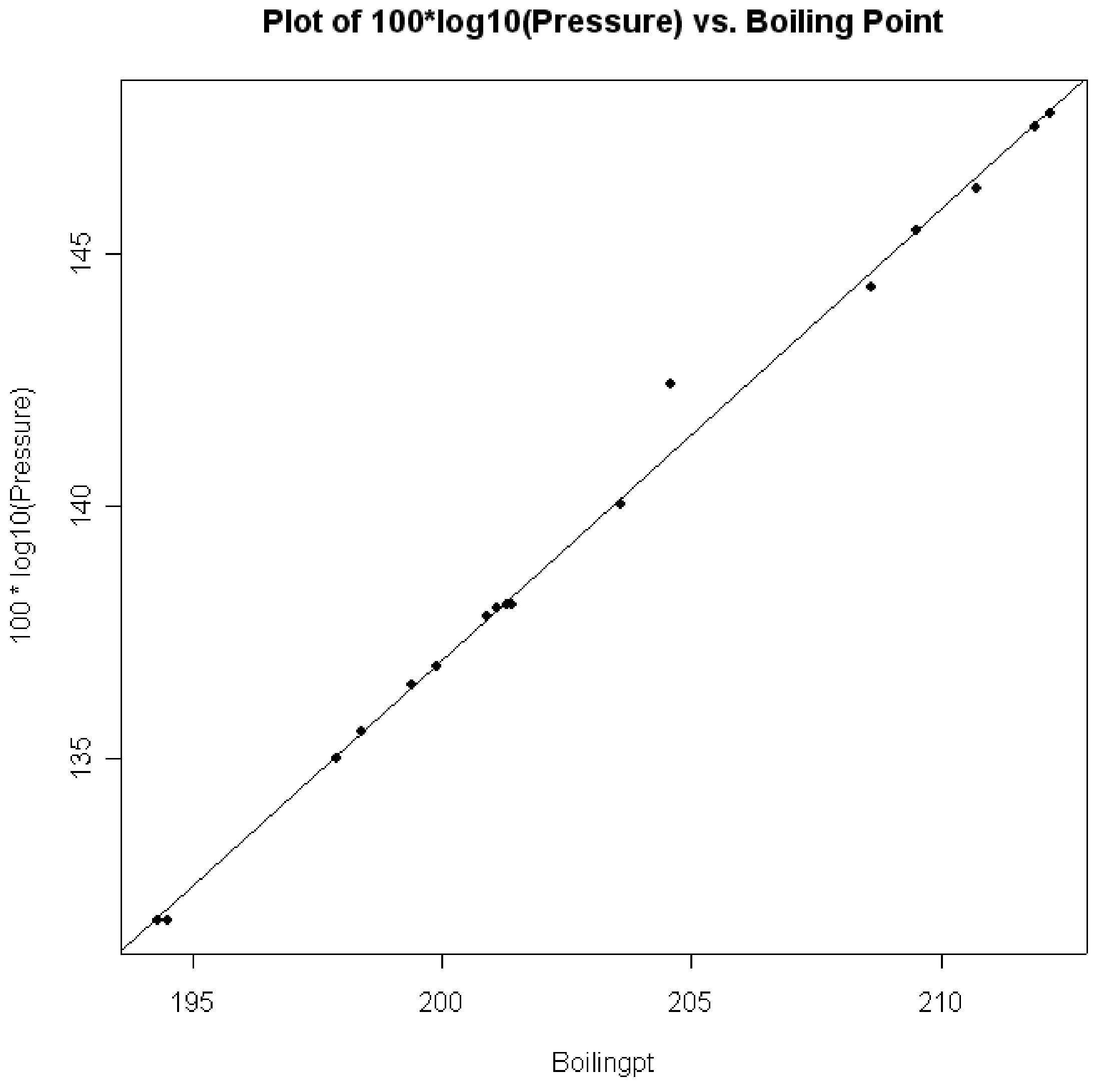
$\hat{\beta}_1 = \dfrac{475.30}{530.78} = 0.8955$ $\hat{\beta}_0 = 139.61 - (0.8955)(202.95) = -42.131$ (Can you verify these results?)

Fitted regression model: 100*log(Pressure) = -42.131 + 0.8955 Boiling Pt

Case	Boiling Pt (^{o}F)	100*Log(Pressure)	Fitted Value	Residual
1	194.5	131.79	132.04	-0.2481
2	194.3	131.79	131.86	-0.0690
3	197.9	135.02	135.09	-0.0538
4	198.4	135.55	135.54	0.0187
5	199.4	136.46	136.43	0.0330
6	199.9	136.83	136.88	-0.0412
7	200.9	137.82	137.77	0.0561
8	201.1	138.00	137.95	0.0584
9	201.4	138.06	138.22	-0.1560
10	201.3	138.04	138.13	-0.0845
11	203.6	140.04	140.19	-0.1471
12	204.6	142.44	141.09	1.3599
13	209.5	145.47	145.48	0.0014
14	208.6	144.34	144.67	-0.3198
15	210.7	146.30	146.55	-0.2429
16	211.9	147.54	147.63	-0.0792
17	212.2	147.80	147.89	-0.0871

Estimate of common variance (σ^2): $\boldsymbol{\hat{\sigma}^2 = \dfrac{RSS}{n-2}} = \dfrac{2.1579}{17-2} = 0.1439$

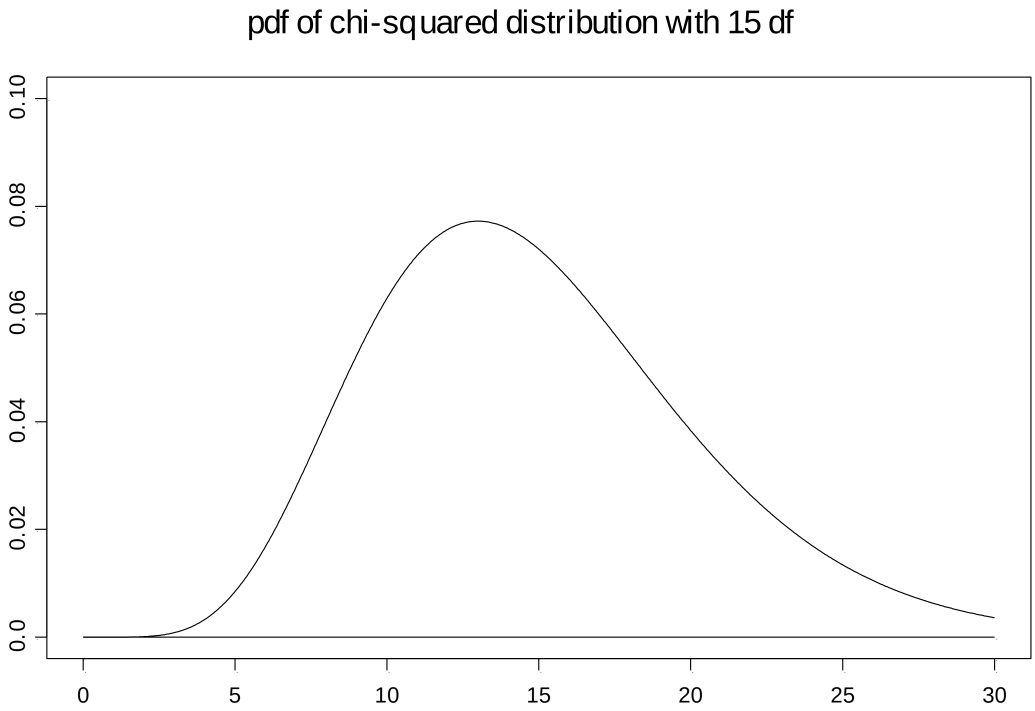
Fitted regression line for Forbes data:



🎥 **Video 1.6**

The χ^2 (chi-squared) distribution

There is a family of χ^2 variables with different degrees of freedom (df), denoted by v. χ^2(v) denotes a chi–squared variable with v df , for v = 1 , 2 , 3,...,



It may be shown that **E(χ^2(v)) = v**, so E(χ^2 (n – 2)) = n – 2 and E(χ^2 (15)) = 15

It may be shown that $\frac{\mathbf{RSS}}{\boldsymbol{\sigma}^2} \sim \boldsymbol{\chi}^2\mathbf{(n-2)}$, so E($\frac{\text{RSS}}{\sigma^2}$) = n – 2 and so E($\frac{\text{RSS}}{n-2}$) = σ^2

Thus $\frac{\text{RSS}}{n-2}$ is an **unbiased** estimate of σ^2

Here n = number of pairs of observations = 17 and RSS = 2.1579.

So and unbiased estimate of σ^2 is $\frac{2.1579}{17-2} = 0.1439$

and $\sqrt{0.1439} = 0.3793$ is an estimate of σ and is called the **standard error of regression** (or the **residual standard error** in R).

Video 1.7

Example of standard error: the sampling distribution of the mean

If $X_1, X_2, \ldots X_n$ are IID with $E(X_i) = \mu$ and var $(X_i) = \sigma^2$,

then $\bar{X} = \frac{1}{n}\sum_{i=1}^{n} X_i$ has mean $E(\bar{X}) = \mu$ and variance $\text{var}(\bar{X}) = \frac{\sigma^2}{n}$

The sample variance $s^2 = \frac{1}{n-1}\sum_{i=1}^{n}(X_i - \bar{X})^2$ is an unbiased estimate of σ^2

Hence $\frac{s^2}{n}$ is an unbiased estimate of $\text{var}(\bar{X}) = \frac{\sigma^2}{n}$

and so $\frac{s}{\sqrt{n}} = \sqrt{\frac{s^2}{n}}$ is an estimate of $\text{sd}(\bar{X}) = \frac{\sigma}{\sqrt{n}}$

$\frac{s}{\sqrt{n}}$ is called the **standard error of the mean** and denoted by $\text{se}(\bar{X})$

In general, the **square root of an estimated variance** is called a **standard error**

Video 1.8

Analysis of Variance

$SYY = \sum_{i=1}^{n}(y_i - \bar{y})^2$ = total sum of squares for the $y_i's$

SYY measures the total variability of the y_i's about their mean

RSS = residual sum of squares = sum of squared residuals

RSS measures that part of the total variability of the y_i's **not explained** by the regression line

Regression Sum of Squares = *SS*reg = *SYY* - *RSS*

*SS*reg measures that part of the total variability of the y_i's **explained** by the regression line

SSreg can be calculated directly from $SSreg = \frac{(SXY)^2}{SXX}$

The fundamental identity of the analysis of variance is

SYY* = *SS*reg + *RSS

Total sum of squares = regression sum of squares + residual sum of squares

For a **perfect fit**, *RSS* = 0 and *SYY* = *SSreg*

The Coefficient of Determination, R^2

The proportion of variability that is explained can be expressed as a percentage of the total variability in the Y's – called the Coefficient of Determination, R^2.

$$R^2 = \frac{\text{SSreg}}{\text{SYY}} = \frac{\text{SYY - RSS}}{\text{SYY}} = \frac{427.76 - 2.1579}{427.76} = 99.50\%$$

Here, 99.50% of the variability in the observed 100*log(pressure) values can be explained by the boiling point.

What would a low value of R^2 mean?

In the case of simple linear regression, $R^2 = r^2$.
$R^2 = (0.9975)^2 = 99.50\%$

🎥 Video 1.9

The F-test for regression

Our regression model is

$$y_i = \beta_0 + \beta_1 x_i + e_i$$

We test the **null hypothesis** $H_0: \beta_1 = 0$
against the **alternative hypothesis** $H_1: \beta_1 \neq 0$

H_0 says that the **true** regression line is **horizontal** ($y_i = \beta_0 + e_i$)
In this case, the least squares estimate of β_0 is $\bar{y}$ and each fitted value is $\hat{y}_i = \bar{y}$ and so
$SYY = \sum(y_i - \bar{y})^2 = \sum(y_i - \hat{y}_i)^2 = RSS$

If H_0 is true, then X is **no use** in predicting Y and X **should not** be included in the regression model.

H_1 says that the true regression line is **not horizontal** and so X is **of use** in predicting Y and **should** be included in the regression model.

Intuitively, the more the fitted line differs from a horizontal line and the more of *SYY* that is explained by the regression, the more we should be inclined to reject H_0 and accept H_1, i.e. to conclude that X is worth including in the regression model.

Here the fitted line is far from horizontal and *SSreg* = 425.61 accounts for most of
SYY = 427.76, which strongly suggests that we should reject H_0 and accept H_1.

If H_0 is true, it may be shown that $\mathbf{E(\,\mathit{SSreg}/1\,) = E(\mathit{RSS}/(n\text{-}2)) = \sigma^2}$

Thus**, if H_0 is true**, the Regression Mean Square is also an **unbiased** estimate of σ^2 and so we would expect the Regression Mean Square and the Residual Mean Square to be about the same size

If H_1 is true, it may be shown that $\mathbf{E(\,\mathit{SSreg}/1\,) > \sigma^2}$, so we would expect the Regression Mean Square to be bigger than the Residual Mean Square

Here *SSreg*/1 = 425.60 while *RSS*/(n-2) = 0.1439.

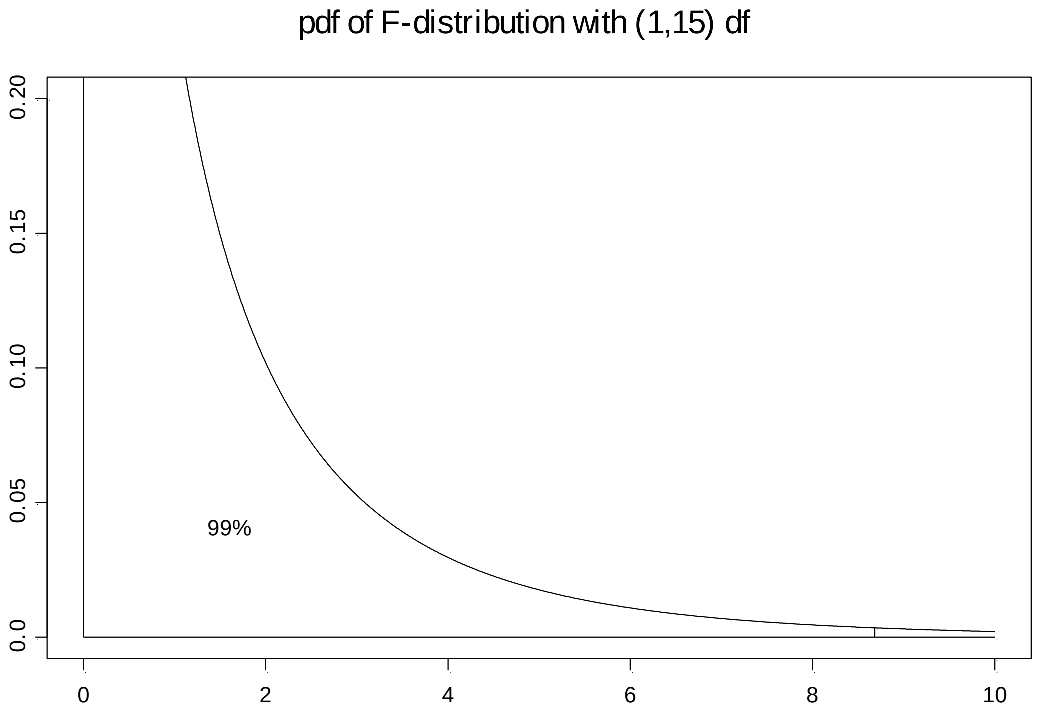
Thus the Regression Mean Square is much bigger than the Residual Mean Square and this provides strong evidence that H_1 is true and H_0 is false

If H_0 is true, it may be shown that

$$F = \frac{SSreg / 1}{RSS / (n-2)} \sim F(1, n-2)$$

If the observed value of *F* is **too big**, we reject H_0 and accept H_1.

The F-distribution
There is a family of *F* variables, each with a pair of degrees of freedom v_1 and v_2.



Critical values of *F* for specified significance levels (e.g. $\alpha = 0.01$) are found in *F*- tables

$v_1 = 1$

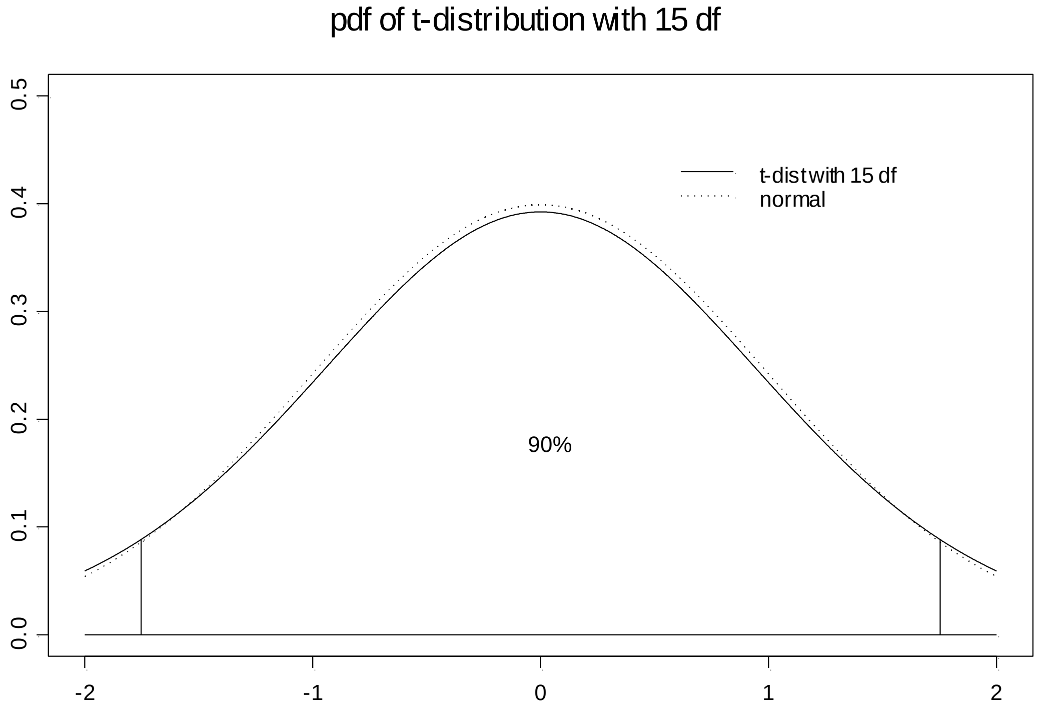
$v_2 = 14$	8.86
$v_2 = 15$	8.68
$v_2 = 16$	8.53

Here $F(0.01;1,15) = 8.68$, while the observed value of *F* is 2958.
Thus is the observed value of *F* is significant at the 1% significance level.
The observed *p*-value << 0.01.

Video 1.10

The t-distribution

There is a family of t-variables with different degrees of freedom (df).
Let t(v) denote a t-variable with v degrees of freedom.
The pdf for t(v) is similar in shape to that of the N(0,1) distribution, but flatter.
As v → ∞, the t(v) → N(0,1).



Critical values for t-distributions are found in t-tables.
$\alpha = 0.05$ v = 15 t = 1.753

Thus 5% of the area under the pdf curve of t(15) lies in the interval (1.753, ∞)
and 5% of the area lies in the interval (∞, -1.753)
and 90% of the area lies in the interval (-1.753,1.753).
Thus each tail of the distribution accounts for 5% of the area under the pdf.

Let $t(\alpha, n-1)$ = critical value of t-distribution with df = n-1 for a two-sided test with significance level α, i.e. an area of $\alpha/2$ in each tail.
t(0.10, 15) = 1.753 t(0.05, 15) = 2.131 t(0.01, 15) = 2.947

t-tests and confidence intervals:

Population mean:

If $X_1, X_2, \ldots X_n$ are NID(μ, σ^2), then $\dfrac{\bar{X} - \mu}{s/\sqrt{n}} = \dfrac{\bar{X} - E(\bar{X})}{se(\bar{X})} \sim t_{n-1}$

The general form in which the t-distribution occurs is: $\dfrac{\text{variable} - E(\text{variable})}{se(\text{variable})} \sim t$

To test the hypothesis H_0: $\mu = \mu_0$ against H_1: $\mu \neq \mu_0$, use the test statistic $t = \dfrac{\bar{X} - \mu_0}{s/\sqrt{n}}$

If the $|t| > t(\alpha, n\text{-}1)$, reject H_0: $\mu = \mu_0$ and accept H_1: $\mu \neq \mu_0$. This test has significance level α.

The general form for a test statistic based on the t-distribution is

$$t = \frac{\text{variable} - E(\text{variable})}{se(\text{variable})}$$

The (1 - α)100% confidence interval for μ is : $\bar{X} \pm t(\alpha, n-1)\dfrac{s}{\sqrt{n}}$

The general form for a (1 - α)100% confidence interval based on the t-distribution is

$$\text{variable} \pm t(\alpha, df)\, se(\text{variable})$$

🎥 Video 1.11

Prediction and fitted values

To predict the actual value of Y at a given value of X, we use the fitted value.
Let se_{pred}= standard error of this prediction.

$se_{pred} = \sigma\sqrt{1 + \frac{1}{n} + \frac{(X_p - \bar{X})^2}{SXX}}$. We replace σ by its estimate sqrt(RSS/(n-2)) to get an estimated standard error.

To estimate the average value of Y at a given value of X, we use the fitted value.
Let se_{fit} = standard error of the fitted value, when used to estimate the average value of Y for given X.

$se_{fit} = \sigma\sqrt{\frac{1}{n} + \frac{(X_p - \bar{X})^2}{SXX}}$. We again replace σ by its estimate sqrt(RSS/(n-2)) to get an estimated standard error.

For the Forbes data, estimate the average Y value (100*log(Pressure)) for a Boiling point of 200.

Fitted regression model: 100*log(Pressure) = -42.131 + 0.8955 Boiling Pt
Fitted value: 100*log(Pressure) = -42.131 + 0.8955 (200) = 136.969

A (1 - α)100% confidence interval for average value of Y at a given value of X is given by: fitted value ± t(α, n-2) se(fitted value)

$$se_{fit} = \hat{\sigma}\sqrt{\frac{1}{n} + \frac{(X_p - \bar{X})^2}{SXX}} = 0.3793\sqrt{\frac{1}{17} + \frac{(200 - 202.95)^2}{530.78}} = 0.1040$$

A 99% confidence interval is 136.969 ± 2.947(0.1040) = (136.663, 137.275)

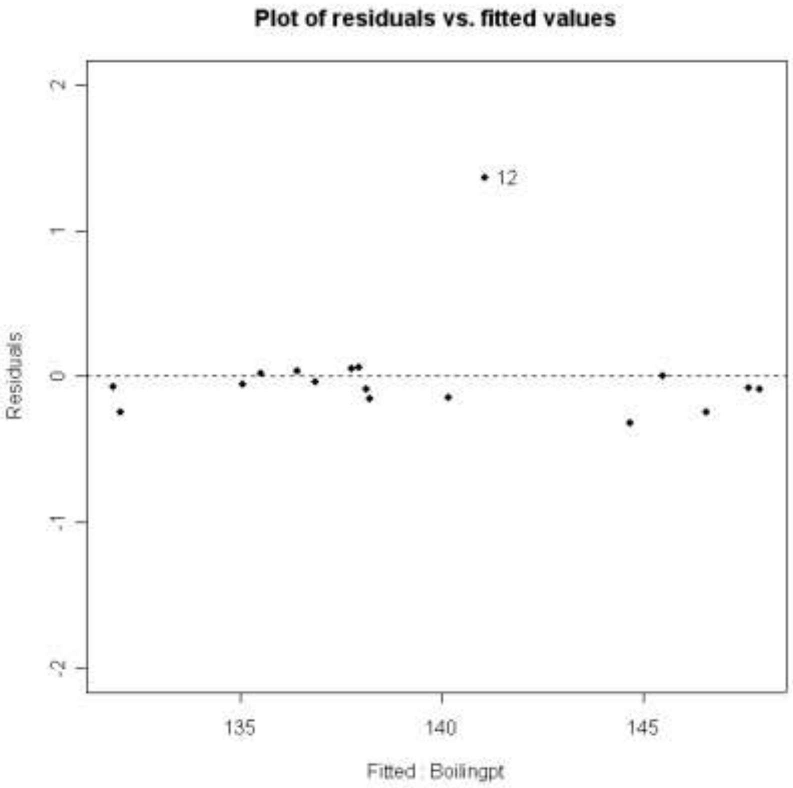
🎥 Video 1.12

Residuals

The residuals can be used to check the assumptions about the error terms and to check the appropriateness of the model.

There are many ways in which the residual are used and these will be the focus of later chapters.

For now, we concentrate on a plot of the residuals versus the fitted values.
Curvature in this plot would suggest that the relationship is not linear and a transformation of one or both variables would be beneficial.
The common variance assumption would be violated if the residuals appeared to increase (or decrease) in magnitude with the predicted values.



There does not appear to be any curvature.

The common variance appears constant.

The residual for case 12 is very large relative to the others.
It might suggest that the assumptions about the errors are not correct.

- Recording/measurement error in case 12? Investigate?
- Impact of including/excluding case 12?

Summer 2006 Question 1

Data were obtained on the lean body mass (the weight without fat in kg) and resting metabolic rate for twelve women who were the subjects in a study of obesity. The scatter plot showed evidence of a linear relationship within the range of the data.

A model of the following form was fitted to these data:

$$Y_i \; = \; \beta_0 \; + \beta_1 X_i + e_i \, , \, e_i \; \sim \text{NID}(\, 0, \sigma^2 \,),$$

where Y = Rate = resting metabolic rate and X = Bodymass = lean body mass.

Excerpts from the R output for this model are shown on the next page.

(a) Identify the estimate of β_1 and interpret it.

(b) Interpret the value of R-Squared.

(c) Test the hypothesis H_0: $\beta_1 = 0$ against H_1: $\beta_1 \neq 0$. Quote the value of the test statistic and the associated p-value. Outline the practical implications of your conclusion.

(d) Find a 95% confidence interval for β_1. Comment on the implications of this interval for part (c).

(e) Find a 95% confidence interval for the mean resting metabolic rate of women with a lean body mass of (i) 50kg and (ii) 75kg. Comment on the appropriateness of the intervals in each case.

R output for Question 1

```
> summary(bodymass.lm)
Call:
lm(formula = Rate ~ Bodymass, data = bodymass.df)

Coefficients:
            Estimate Std. Error t value Pr(>|t|)
(Intercept)  201.162    181.701   1.107 0.294169
Bodymass      24.026      4.174   5.756 0.000184 ***
---
Signif. codes:  0 `***' 0.001 `**' 0.01 `*' 0.05 `.' 0.1 `
' 1

Residual standard error: 95.08 on 10 degrees of freedom
Multiple R-Squared: 0.7682,     Adjusted R-squared: 0.745
F-statistic: 33.13 on 1 and 10 DF,  p-value: 0.0001836

> qt(0.975,10)
[1] 2.228139

> predict(bodymass.lm,
data.frame(Bodymass=c(50,75)),se.fit=T,
+ interval="confidence", level = 0.95)
$fit
       fit      lwr      upr
1 1402.465 1313.370 1491.560
2 2003.117 1699.602 2306.631

$se.fit
        1         2
 39.98617 136.21886

> summary(bodymass.df)
    Bodymass          Rate
 Min.   :33.10   Min.   : 913
 1st Qu.:39.25   1st Qu.:1106
 Median :42.00   Median :1230
 Mean   :43.03   Mean   :1235
 3rd Qu.:49.02   3rd Qu.:1402
 Max.   :54.60   Max.   :1502
```

Practical (Assignment) 1

Instructions for this practical

- Open the template “Surname Forename Chpt x” from Canvas (in “Practicals”).
- Complete the grid on the first page.
- Save this file (as a Word document) using your own surname, forename and the appropriate chapter number.

- Practice Question:
- Type the commands one by one into R.
- Compare the results in the R text output and graphics with the corresponding results and figures in your notes.
- Use appropriate R output to answer the questions, adapting the R code if necessary.

- Exam Question:
- Adapt the relevant R code you used for the practice question to answer the questions.
- Copy and paste the relevant R text output and graphics into your Word document to support your answers. Change the text font to “Courier New” to align columns.

- Restrict your Word document to a **maximum of 2 pages** (re-sizing graphics and deleting irrelevant R output will help).
- Submit this Word document **via Canvas** by **5.00pm 16th October 2020** (**STRICT** deadline)
- Note that submitting the practical is a declaration that the practical is your own work. Plagiarism/copying will not be tolerated.

Practice Question (not to be submitted)

Data on the Boiling Point of water (°F) and Atmospheric Pressure (in. Hg) were recorded at 17 locations in the **forbes.txt** dataset. Fit a model of the following form to this data:

$$Y_i = \beta_0 + \beta_1 X_i + \varepsilon_i , \varepsilon_i \sim IN(0,\sigma^2) ,$$

where Y = 100 log(Pressure) and X = Boiling Point.

(a) Identify the estimate of β_1 and interpret it.

(b) Interpret the value of R-Squared.

(c) Test the hypothesis H_0: $\beta_1 = 0$ against H_1: $\beta_1 \neq 0$.
Outline the practical implications of your conclusion.

(d) Find 99% confidence intervals for β_0 and β_1.

(e) Find a 99% confidence interval for the mean value of Y for X = 200.

(f) Find a 99% prediction interval for the value of Y for X = 200.

Exam Question (Winter 2019-20, Question 1) (to be submitted)

A recruitment consulting company has taken a random sample of executives running private companies to investigate the variables that potentially influence the salaries of those executives. The data are stored in **Executives.txt (on Canvas)**.

Fit a model of the following form to these data:

$$Y_i = \beta_0 + \beta_1 X_i + e_i, \quad e_i \sim NID(0, \sigma^2),$$

where
Y = Salary = salary (€) of the executive,
X = Experience = experience (years) of the executive.

(a) Prepare an appropriate scatter-plot for these data. Does this scatter-plot suggest that it would be appropriate to fit a linear regression model to these data? Explain. (9 marks)

(b) Interpret the values of the intercept and slope. (8 marks)

(c) Calculate a 98% confidence interval for the slope. Interpret this confidence interval. (10 marks)

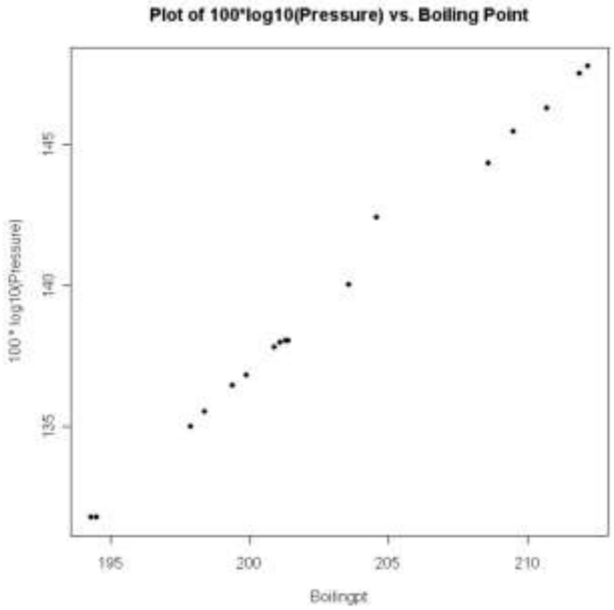
(d) Use the confidence interval in part (c) to test the hypothesis H_0: $\beta_1 = 0$ against H_1: $\beta_1 \neq 0$. Outline the practical implications of your conclusion. Comment of the level of significance used. (15 marks)

(e) Interpret the value of *R*-Squared. What recommendation would you make for the current model? (8 marks)

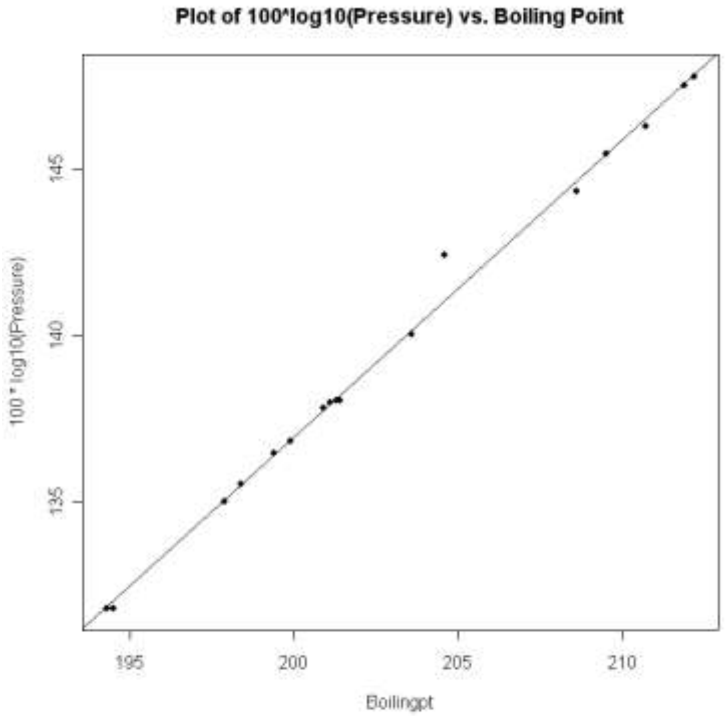
# R code and output for Chapter 1

```
> # read in Forbes data
> forbes.df <-
+
read.table("P:\\ST2053\\forbes.txt",header=T
)
> forbes.df
   Boilingpt Pressure
1      194.5    20.79
2      194.3    20.79
3      197.9    22.40
4      198.4    22.67
5      199.4    23.15
6      199.9    23.35
7      200.9    23.89
8      201.1    23.99
9      201.4    24.02
10     201.3    24.01
11     203.6    25.14
12     204.6    26.57
13     209.5    28.49
14     208.6    27.76
15     210.7    29.04
16     211.9    29.88
17     212.2    30.06

> attach(forbes.df)

> # Scatter plot
> plot(Boilingpt, 100*log10(Pressure),
+ main="Plot of 100*log10(Pressure) vs. Boiling Point",
+ pch=16)
```

Note: you will need to change P:\\ST2053\\ to the appropriate path on your own computer at which you saved the forbes.txt file from Canvas.
This applies to all such paths throughout these notes.



```
> # Fit regression line
> forbes.lm <- lm(100*log10(Pressure)~ Boilingpt,
+ data=forbes.df)

> coef(forbes.lm)
(Intercept)   Boilingpt
-42.1641838   0.8956178

> # Scatter plot with fitted line
> plot(Boilingpt, 100*log10(Pressure),
+ main="Plot of 100*log10(Pressure) vs. Boiling Point",
+ pch=16)

> abline(forbes.lm)
```

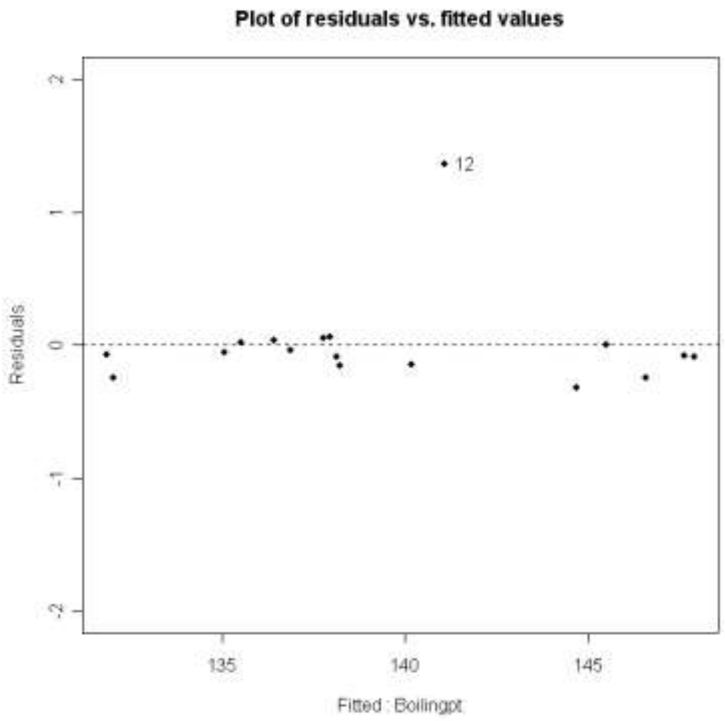

Plot of 100^log10(Pressure) vs. Boiling Point
100 * log10(Pressure)
135
140
145
195
200
205
210
Boilingpt

```
> # Residual standard error
> # Multiple R-squared
> # Standard errors of regression coefficients
> summary(forbes.lm)

Call:
lm(formula = 100 * log10(Pressure) ~ Boilingpt, data =
forbes.df)

Residuals:
     Min       1Q   Median       3Q      Max
-0.31974 -0.14707 -0.06890  0.01877  1.35994

Coefficients:
             Estimate Std. Error t value Pr(>|t|)
(Intercept) -42.16418    3.34136  -12.62 2.17e-09 ***
Boilingpt     0.89562    0.01646   54.42  < 2e-16 ***
---
Signif. codes:  0 `***' 0.001 `**' 0.01 `*' 0.05 `.' 0.1 `
' 1

Residual standard error: 0.3792 on 15 degrees of freedom
Multiple R-Squared: 0.995,      Adjusted R-squared: 0.9946
F-statistic:  2962 on 1 and 15 DF,  p-value: < 2.2e-16

> # critical value of t-distribution
> qt(0.95,15)
[1] 1.753050

> # 90% confidence limits for beta-0
> -42.16418 + (qt(0.95,15)*3.34136)
[1] -36.30661
> -42.16418 - (qt(0.95,15)*3.34136)
[1] -48.02175

> # 90% confidence limits for beta-1
> 0.89562 + (qt(0.95,15)*0.01646)
[1] 0.9244752
> 0.89562 - (qt(0.95,15)*0.01646)
[1] 0.8667648
```

```
> # Anova table, F-test
> anova(forbes.lm)
Analysis of Variance Table

Response: 100 * log10(Pressure)
          Df Sum Sq Mean Sq F value    Pr(>F)
Boilingpt  1 425.76  425.76  2961.5 < 2.2e-16 ***
Residuals 15   2.16    0.14
---
Signif. codes:  0 `***' 0.001 `**' 0.01 `*' 0.05 `.' 0.1 `
' 1

> # 99% confidence interval for the mean of y
> predict(forbes.lm, data.frame(Boilingpt=200),se.fit=T,
+ interval="confidence",level=0.99)
$fit
          fit      lwr      upr
[1,] 136.9594 136.6529 137.2659

$se.fit
[1] 0.1040112

$df
[1] 15

$residual.scale
[1] 0.3791592

> # critical value of t-distribution
> qt(0.995,15)
[1] 2.946713

> # alternative derivation of
> # 99% confidence interval for the mean of y
> 136.9594 + (qt(0.995,15)*0.104)
[1] 137.2659
> 136.9594 - (qt(0.995,15)*0.104)
[1] 136.6529
```

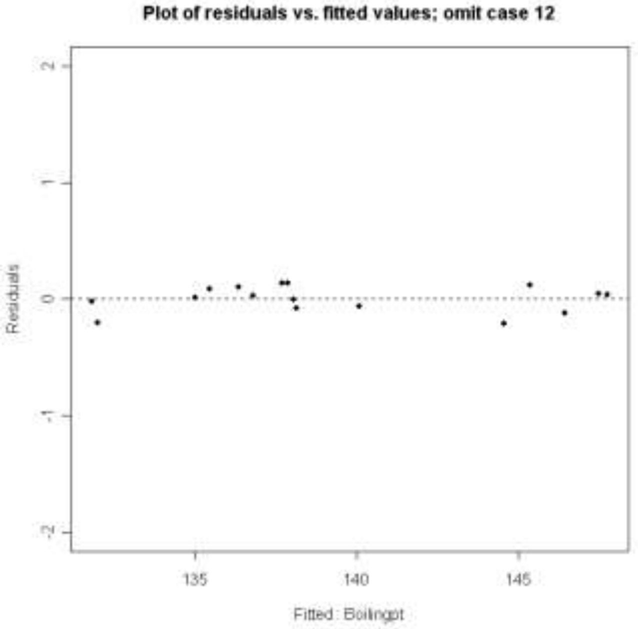
```
> # 99% prediction interval for individual value of y
> predict(forbes.lm, data.frame(Boilingpt=200),se.fit=T,
+ interval="prediction",level=0.99)
$fit
          fit      lwr      upr
[1,] 136.9594 135.8008 138.1179

$se.fit
[1] 0.1040112

$df
[1] 15

$residual.scale
[1] 0.3791592
```

```
> # Plot of residuals vs. fitted values
> plot(fitted(forbes.lm),resid(forbes.lm),
+ main="Plot of residuals vs. fitted values",
+ xlab="Fitted : Boilingpt",ylab="Residuals",
+ ylim=c(-2,2),pch=16)

> abline(h=0,lty=2)
```



```
> # Omitting a case
> # identify outlier
> outlier <-
identify(fitted(forbes.lm),resid(forbes.lm),n=1)
> outlier
[1] 12
> subset.forbes.lm <- lm(100*log10(Pressure)~ Boilingpt,
+ data=forbes.df, subset = -outlier)

> summary(subset.forbes.lm )
```

```
Call:
lm(formula = 100 * log10(Pressure) ~ Boilingpt, data =
forbes.df,
    subset = -outlier)

Residuals:
     Min       1Q   Median       3Q      Max
-0.20882 -0.06338  0.01974  0.08842  0.13558

Coefficients:
              Estimate Std. Error t value Pr(>|t|)
(Intercept) -41.334683   1.003312   -41.2 5.16e-16 ***
Boilingpt     0.891110   0.004944   180.2  < 2e-16 ***
---
Signif. codes:  0 `***' 0.001 `**' 0.01 `*' 0.05 `.' 0.1 `
' 1

Residual standard error: 0.1136 on 14 degrees of freedom
Multiple R-Squared: 0.9996,     Adjusted R-squared: 0.9995
F-statistic: 3.249e+04 on 1 and 14 DF,  p-value: < 2.2e-16

> # Plot of residuals vs.fitted values
> plot(fitted(subset.forbes.lm),resid(subset.forbes.lm),
+ main="Plot of residuals vs. fitted values; omit case 12",
+ xlab="Fitted : Boilingpt",ylab="Residuals",ylim=c(-2,2),
+ pch=16)
> abline(h=0,lty=2)
```



```
> # quit R
> q("yes")
```