23 Regression analysis: method of least squares

The main purpose of regression is to explore the dependence of one variable (Y) on another variable (X).

23.1 Introduction (P.532-555)

Suppose we are interested in studying the relation between two variables X and Y, where X is regarded as an *explanatory (controlled; independent)* variable that may be measured *without error*, and Y is a *response (outcome; dependent)* variable.

We may have

$$Y = \alpha + \beta X,$$

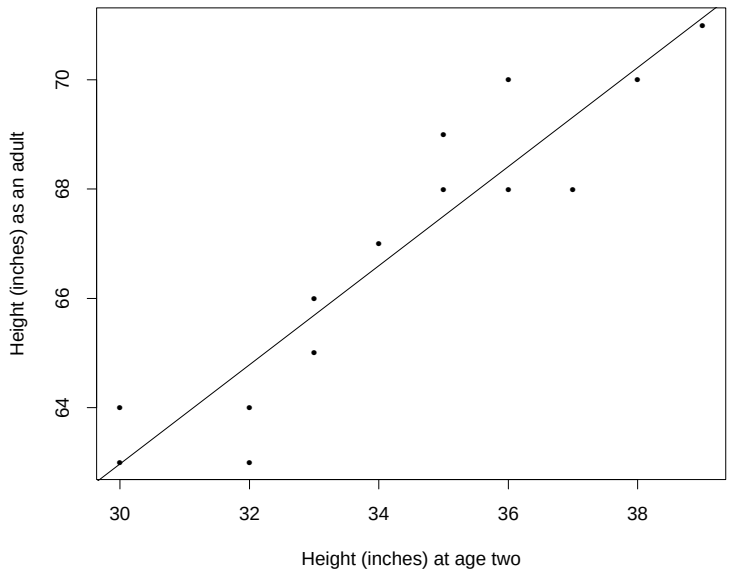
or more generally, $Y = f(X)$. This is called *deterministic* mathematical relation because it does not allow for any error in predicting Y as a function of X.

In this course, we are interested in studying the relationship between X and Y which subject to random errors.

Example: (Predicting height) Parents are often interested in predicting the eventual heights of their children. The following is a portion of the data taken from a study of heights of boys.

```
Height (inches)     39 30 32 34 35 36 36 30
at age two (x_i)    33 37 33 38 32 35

Height (inches)     71 63 63 67 68 68 70 64
as an adult (y_i)   65 68 66 70 64 69
```

A scatter plot of the data is given as follows:



It is clear from the figure that the expected value of Y increases as a linear function of X but with errors. Hence a deterministic relation is NOT suitable.

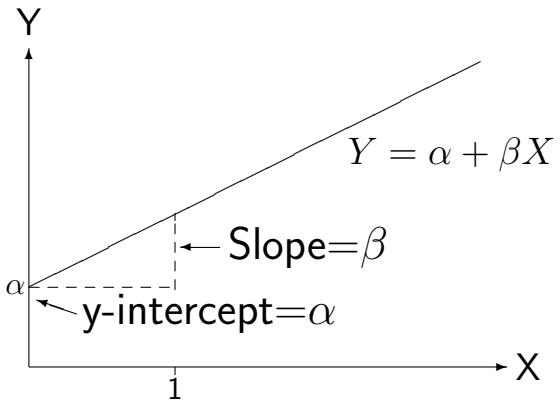
23.2 Linear regression

With repeated experiments, one would find that values of Y often vary in a random manner. Hence the *probabilistic* model

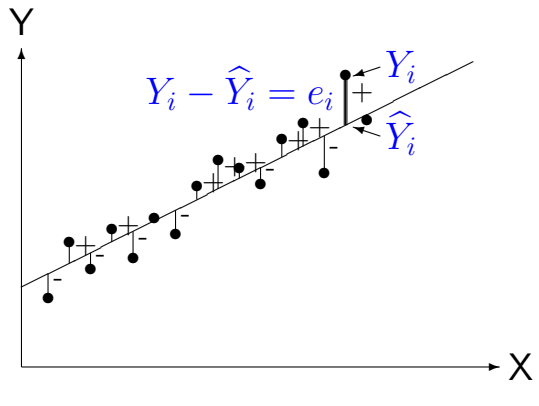
$$Y = \alpha + \beta X + \epsilon$$

where ϵ is a random variable which follows a probability distribution with mean zero provides a more accurate description of the relationship between X and Y in most cases.

The model is generally called a (simple linear) *regression* model or the regression of Y on X. The variables α, β are called *parameters* and they are the *intercept* and *slope* (coefficient) of the linear regression model respectively.



The regression line



Residuals

In practice, the regression function $f(x)$ may be more complex than a simple linear function. For example, we may have

$$\begin{aligned} f(x) &= \alpha + \beta_1 x + \beta_2 x^2, \quad \text{or} \\ f(x) &= \alpha + \beta x^{\frac{1}{2}}, \quad \text{etc.} \end{aligned}$$

When we say we have a *linear regression* for Y, we mean that $f(x)$ is a linear function of the unknown parameters α, β, etc, not necessarily a linear function of x.

23.3 Fitting a straight line by least squares (P.555-558)

Suppose we have a set of observed bivariate data

$$(x_1, y_1),\ (x_2, y_2),\ \ldots,\ (x_n, y_n).$$

A scatter plot is drawn and indicates that a straight line model is appropriate. We need to choose a line (regression line)

$$y = \hat{\alpha} + \hat{\beta}x,$$

which fit the data most closely. In general, we call $y = \hat{\alpha} + \hat{\beta}x$ the *regression line* or *fitted line*.

A lot of techniques, such as *least squares*, *robust methods*, etc, can be used to produce a reasonable estimates $\hat{\alpha}, \hat{\beta}$ for the unknown regression function $f(x)$.

Note that the *expected* value or the *mean* of Y_i given $X = x_i$ under the model is

$$E(Y_i|X = x_i) = \hat{y}_i = \alpha + \beta\, x_i$$

and the errors or residuals of the model when $Y_i = y_i$ is

$$\epsilon_i = y_i - \hat{y}_i = y_i - \alpha - \beta x_i.$$

By the method of *least squares*, the values $\hat{\alpha}$ and $\hat{\beta}$ are estimated such that the sum of squared residuals (SSR) is a minimum. That is

$$S(\hat{\alpha}, \hat{\beta}) \ = \ \min_{\alpha,\beta} S(\alpha, \beta),$$

where the objective function is

$$S(\alpha, \beta) = SSR = \sum_{i=1}^{n} \epsilon_i^2 = \sum_{i=1}^{n} (y_i - \alpha - \beta\, x_i)^2.$$

Then we solve $\dfrac{\partial SSR}{\partial \alpha} = 0$ and $\dfrac{\partial SSR}{\partial \beta} = 0$ for $\hat{\alpha}$ & $\hat{\beta}$.

The solutions are

$$\hat{\beta} = \frac{S_{xy}}{S_{xx}} = \frac{\sum_{i=1}^{n}(x_i - \bar{x})(Y_i - \bar{Y})}{\sum_{i=1}^{n}(x_i - \bar{x})^2} = \frac{\sum_{i=1}^{n}(x_i - \bar{x})Y_i}{\sum_{i=1}^{n}(x_i - \bar{x})^2}$$

and

$$\hat{\alpha} = \bar{Y} - \hat{\beta}\bar{x}.$$

Note:

$$\sum_{i=1}^{n}(x_i - \bar{x})\bar{Y} = \bar{Y}\sum_{i=1}^{n}(x_i - \bar{x}) = \bar{Y}(n\bar{x} - n\bar{x}) = 0,$$

and $\hat{\alpha}$ and $\hat{\beta}$ are random variables as they both depend on ϵ_i through the Y_i. The x_i are fixed numbers.

23.4 Calculations of $\hat{\alpha}$ and $\hat{\beta}$

Example: (Predicting height)

Note that

$$\bar{x} = 34.286, \quad \bar{y} = 66.857,$$
$$\sum_{i=1}^{n} x_i^2 = 16558, \quad \sum_{i=1}^{n} y_i^2 = 62674,$$
$$\sum_{i=1}^{n} x_i y_i = 32183.$$

Solution:

$$\begin{aligned} S_{xx} &= \sum_{i=1}^{n} x_i^2 - n(\bar{x})^2 = 16558 - 14(34.286^2) = 100.86, \\ S_{xy} &= \sum_{i=1}^{n} x_i y_i - n(\bar{x}\bar{y}) = 32183 - 14(34.286)(66.857) = 91.57, \\ \hat{\beta} &= \frac{S_{xy}}{S_{xx}} = \frac{91.57}{100.86} = 0.9079, \\ \hat{\alpha} &= \bar{y} - \hat{\beta}\bar{x} = 66.857 - 0.9079(34.286) = 35.73 \end{aligned}$$

Hence the fitted least squares line is

$$\hat{y} = \hat{\alpha} + \hat{\beta}\, x = 35.73 + 0.9079x.$$

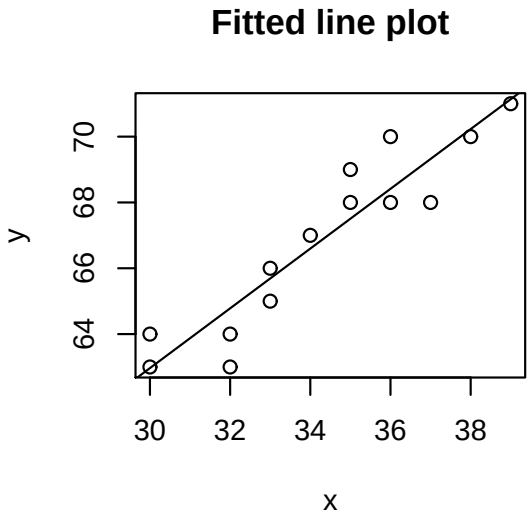
The R code to compute the least squares line including the residuals of the fit is

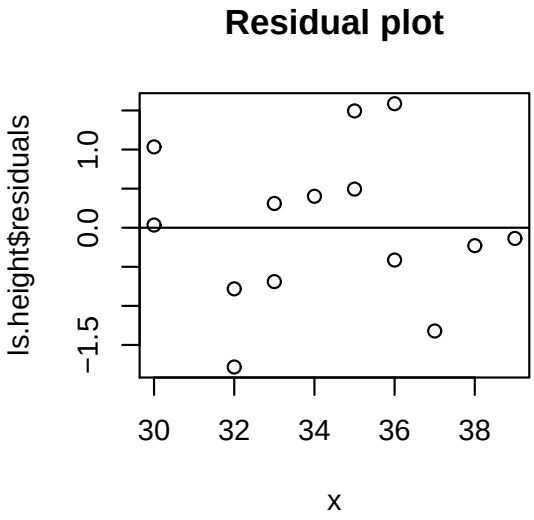
`lsfit(x,y)`

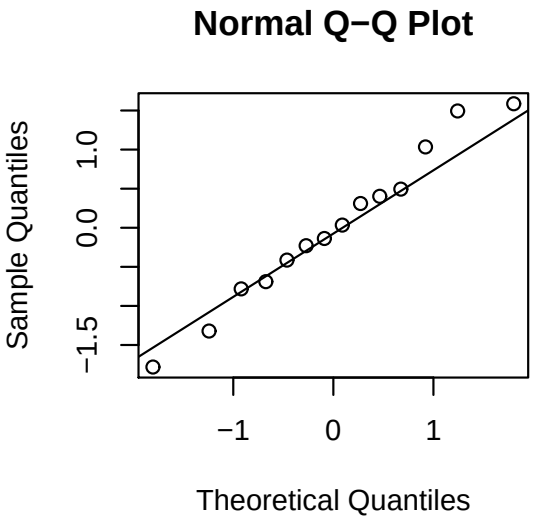
```
> x=c(39,30,32,34,35,36,36,30,33,37,33,38,32,35)
> y=c(71,63,63,67,68,68,70,64,65,68,66,70,64,69)
```

```
> ls.height=lsfit(x,y)
> ls.height$coef
Intercept         X
35.728045  0.907932

> ls.height$residuals
 [1] -0.13739377  0.03399433 -1.78186969  0.40226629
      0.49433428 -0.41359773  1.58640227  1.03399433
 [9] -0.68980170 -1.32152975  0.31019830 -0.22946176
     -0.78186969  1.49433428

> par(mfrow=c(2,2))   #plots
> plot(x,y)
> abline(ls.height)
> title("Fitted line plot")
> plot(x,ls.height$resid)
> abline(h=0)
> title("Residual plot")
> qqnorm(ls.height$resid)
> qqline(ls.height$resid)

> xm=mean(x)   #check
> ym=mean(y)
> c(xm,ym)
[1] 34.28571   66.85714
> Sxx=sum(x^2)-sum(x)^2/n
> Sxy=sum(x*y)-sum(x)*sum(y)/n
> c(Sxx,Sxy)
[1] 100.8571   91.57143
```

```
> beta=Sxy/Sxx
> alpha=mean(y)-beta*mean(x)
> c(alpha,beta)
[1] 35.72805   0.907932
```







23.5 Checking residuals

The regression model above does not specify the distribution for Y_i or ϵ_i. In fact, we usually assume that

$$Y_i \sim \mathcal{N}(\alpha + \beta x_i, \sigma^2)$$

or equivalently, $\epsilon_i \overset{\text{iid}}{\sim} \mathcal{N}(0, \sigma^2)$ where 'iid' stands for 'independent and identically distributed'.

Model assumptions:

1. *Linearity* of data, i.e. $y_i = \alpha + \beta x_i + \epsilon_i$,
2. *Equality* of variance, i.e. a common σ^2 independent of x_i and
3. *Normality* of residuals, i.e. $\epsilon_i \sim \mathcal{N}(0, \sigma^2)$.

These assumptions may be checked by using

1. Linearity of data: the fitted line plot of y_i,
2. Equality of variance: the scatter plot of residuals $\epsilon_i = y_i - \hat{y}_i$ against the fitted values $\hat{y}_i$,
3. Normality of residuals: the normal `qq`-plot of the residuals ϵ_i.

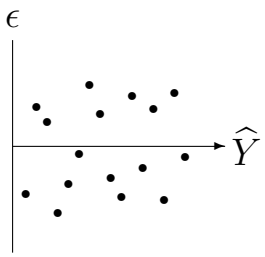
The residual plot should be a *random scatter* around 0 with no particular pattern.

Violation to assumptions on residuals:

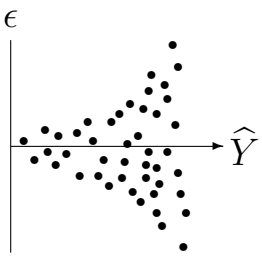
1. If $E(\epsilon_i) = a$ say, we can just increase α increase by a and all ϵ_i will drop by a. Then $E(\epsilon_i) = 0$.
2. If $\sigma^2 = \sigma_i^2$ depends on x_i, e.g. the variability of the stock market index usually increases with time, we may consider models that allow σ_i^2 to change across x_i.
3. If ϵ_i does not follow a normal distribution, we may consider other distributions say t with wider spread.

4. If ϵ_i and ϵ_j for some $i \neq j$ are dependent, we may consider models that allow a very general variance-covariance structure.

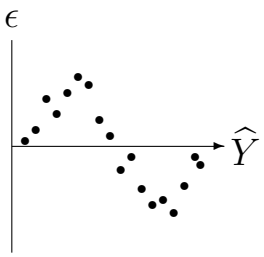
A random scatter plot indicates the model assumptions are OK.



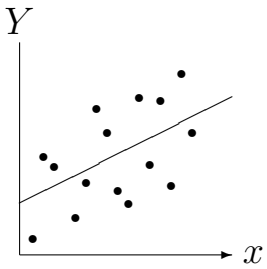
Nonconstant variance σ^2 which increases with x.

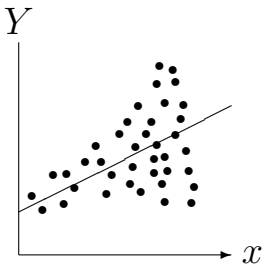


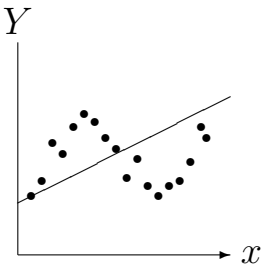
Functional form $f(x)$ may be wrong. It should be say $f(x) = \alpha + \beta_1 x + \beta_2 x^2 + \beta_3 x^3$.



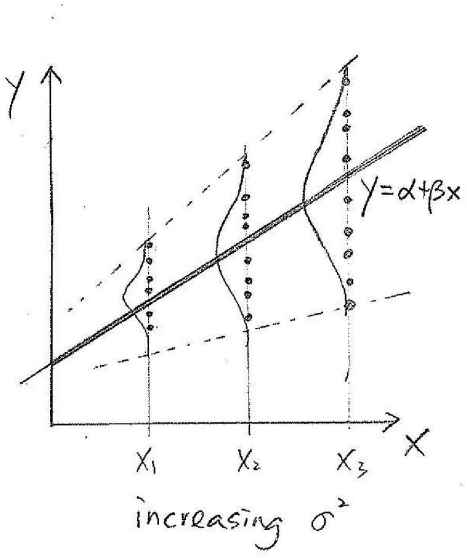
Residuals plots ϵ vs $\widehat{Y}$

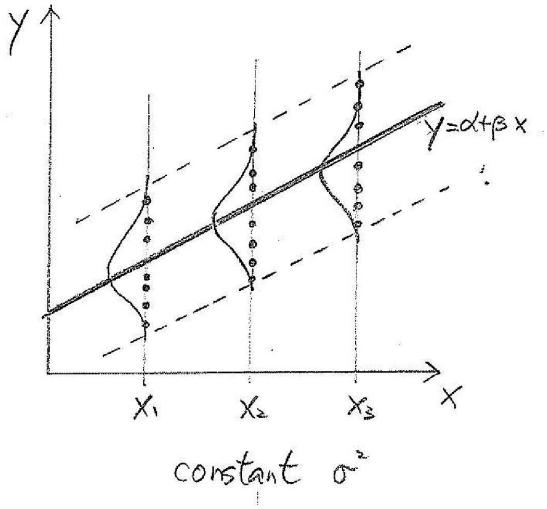






Fitted line Y vs x





The least square regression line is sensitive to model assumptions and outliers. If there are pattens in the residuals, one needs to modify the fitted model using transformed data.

24 Regression analysis: distribution of parameters

24.1 Properties of least squares estimates (P.558-570)

Result 1. The intercept $\hat{\alpha}$ and slope $\hat{\beta}$ provide *unbiased* estimates of the true intercept α and slope β, i.e.

$$E(\hat{\alpha}) = \alpha \quad \text{and} \quad E(\hat{\beta}) = \beta.$$

Proof:

$$\begin{aligned}
E(\hat{\beta}) &= E\Big[\frac{\sum_{i=1}^{n}(x_i-\bar{x})(Y_i-\bar{Y})}{S_{xx}}\Big] = \frac{\sum_{i=1}^{n}(x_i-\bar{x})E(Y_i) - E(\bar{Y})\sum_{i=1}^{n}(x_i-\bar{x})}{S_{xx}} \\
&= \frac{\sum_{i=1}^{n}(x_i-\bar{x})(\alpha+\beta x_i)}{S_{xx}} = \frac{\alpha\sum_{i=1}^{n}(x_i-\bar{x})}{S_{xx}} + \frac{\beta\sum_{i=1}^{n}(x_i-\bar{x})x_i}{S_{xx}} \\
&= \frac{\beta\sum_{i=1}^{n}(x_i-\bar{x})(x_i-\bar{x})}{S_{xx}} = \frac{\beta S_{xx}}{S_{xx}} = \beta \quad (\text{since } \bar{x}\sum_{i=1}^{n}(x_i-\bar{x}) = 0) \\
E(\hat{\alpha}) &= E(\bar{Y}) - E(\hat{\beta})\bar{x} = \frac{1}{n}\sum_{i=1}^{n}E(Y_i) - \beta\bar{x} = \frac{1}{n}\sum_{i=1}^{n}(\alpha+\beta x_i) - \beta\bar{x} \\
&= \alpha + \beta\frac{1}{n}\sum_{i=1}^{n}x_i - \beta\bar{x} = \alpha + \beta\bar{x} - \beta\bar{x} = \alpha
\end{aligned}$$

since x_i are fixed, $\sum_{i=1}^{n}(x_i-\bar{x}) = 0$ and $E(Y_i) = \alpha+\beta x_i+E(\epsilon_i) = \alpha+\beta x_i$.

Result 2. Under the assumptions of the regression model, we have

$$\begin{aligned} Var(\hat{\alpha}) &= \sigma^2\Big(\frac{1}{n}+\frac{(\bar{x})^2}{S_{xx}}\Big) = \frac{\sigma^2\sum_i x_i^2}{n\,S_{xx}} \\ Var(\hat{\beta}) &= \sigma^2/S_{xx}, \\ Cov(\hat{\alpha},\,\hat{\beta}) &= -\frac{\bar{x}\,\sigma^2}{S_{xx}}. \end{aligned}$$

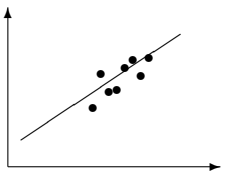
Proof:

$$\begin{aligned} Var(\hat{\beta}) &= Var\Big(\frac{\sum_{i=1}^{n}(x_i-\bar{x})Y_i}{S_{xx}}\Big) = \frac{\sum_{i=1}^{n}(x_i-\bar{x})^2 Var(Y_i)}{S_{xx}^2} \\ &= \frac{S_{xx}\sigma^2}{S_{xx}^2} = \frac{\sigma^2}{S_{xx}} \quad \text{since } Var(aY) = a^2 Var(Y) \\ Var(\hat{\alpha}) &= Var(\bar{Y}-\hat{\beta}\bar{x}) = Var(\bar{Y}) + \bar{x}^2 Var(\hat{\beta}) = \frac{\sigma^2}{n}+\frac{\bar{x}^2\sigma^2}{S_{xx}} \\ &= \sigma^2\Big(\frac{1}{n}+\frac{\bar{x}^2}{S_{xx}}\Big) = \frac{\sigma^2(S_{xx}+n\bar{x}^2)}{nS_{xx}} \\ &= \frac{\sigma^2(\sum_{i=1}^{n}x_i^2 - n\bar{x}^2 + n\bar{x}^2)}{nS_{xx}} = \frac{\sigma^2\sum_{i=1}^{n}x_i^2}{nS_{xx}} \end{aligned}$$

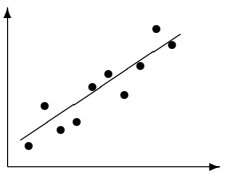
since $Y_1, \ldots, Y_n$ and β are independent. The proof for $Cov(\hat{\alpha}, \hat{\beta})$ is more complicated and hence is omitted. Note that both variances are divided by S_{xx}.

Hence $Var(\hat{\beta})$ increases with σ^2, the variability of residuals ϵ_i, and decreases with S_{xx} which measures the variability of x_i.

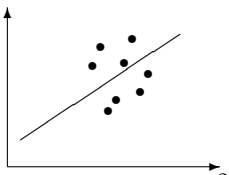
The fit of a model depends on how big is SSR relative to the total spread of the data as measured by S_{xx}.



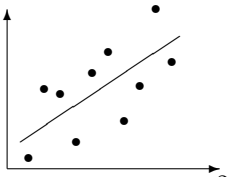
Small SSR, σ^2
Small total spread S_{xx}



Small SSR, σ^2
Large total spread S_{xx}
The best $\widehat{\beta}$



Large SSR, σ^2
Small total spread S_{xx}
The worst $\widehat{\beta}$



Large SSR, σ^2
Large total spread S_{xx}

Good fit and bad fit

25 Regression analysis: estimation theory and inference

25.1 Estimate of σ^2 (P.691-693)

For the simple linear regression model,

$$Y_i = \alpha + \beta\, x_i + \epsilon_i, \quad \epsilon_i \overset{\text{iid}}{\sim} \mathcal{N}(0, \sigma^2)$$

The variances of $\hat{\alpha}$ and $\hat{\beta}$ depend on the residuals variance σ^2. Hence we need to estimate σ^2.

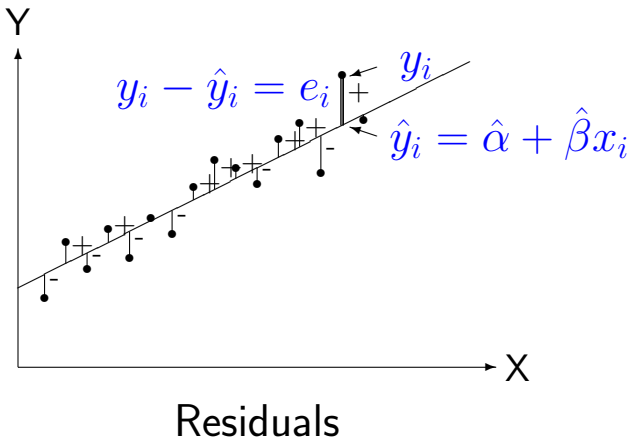
Since σ^2 is the variance of residuals, an estimate S^2 of σ^2 is based on the average of the squared residuals called *Residual Sum of Squares* (SSR)

$$SSR = \sum_{i=1}^{n} \epsilon_i^2 = \sum_{i=1}^{n} (y_i - \hat{\alpha} - \hat{\beta}\, x_i)^2.$$

It can be shown that

$$S^2 = \frac{SSR}{n-2}$$

is an unbiased estimate of σ^2, i.e., $E(S^2) = \sigma^2$.



Residuals

Note that

$$\begin{aligned} SSR &= \sum_{i=1}^{n}(y_i - \hat{\alpha} - \hat{\beta}\, x_i)^2 = \sum_{i=1}^{n}[(y_i - \bar{y}) - \hat{\beta}(x_i - \bar{x})]^2 \\ &= \sum_{i=1}^{n}(y_i - \bar{y})^2 - 2\hat{\beta}\sum_{i=1}^{n}(x_i - \bar{x})(y_i - \bar{y}) + \hat{\beta}^2\sum_{i=1}^{n}(x_i - \bar{x})^2 \\ &= S_{yy} - 2\hat{\beta}S_{xy} + \hat{\beta}\frac{S_{xy}}{S_{xx}}S_{xx} \\ &= S_{yy} - \hat{\beta}\, S_{xy} = SST_o - SST \\ \overset{\text{or}}{=}\ & S_{yy} - \frac{S_{xy}}{S_{xx}}\, S_{xy} = \frac{S_{yy}S_{xx} - S_{xy}^2}{S_{xx}} \end{aligned}$$

since $\hat{\alpha} = \bar{y} - \hat{\beta}\bar{x}$ and $\hat{\beta} = \dfrac{S_{xy}}{S_{xx}}$.

25.2 Distributions of the estimators $\hat{\alpha}$, $\hat{\beta}$ and S^2

Assume that we have a linear regression model:

$$Y_i = \alpha + \beta\, x_i + \epsilon_i, \quad \epsilon_i \overset{\text{iid}}{\sim} \mathcal{N}(0, \sigma^2), \quad i = 1, 2, ..., n.$$

Under this model,

$$\frac{\hat{\alpha} - \alpha}{\sqrt{Var(\hat{\alpha})}} = \frac{\hat{\alpha} - \alpha}{\sigma\sqrt{\frac{1}{n} + \frac{(\bar{x})^2}{S_{xx}}}} \sim \mathcal{N}(0, 1),$$

$$\frac{\hat{\beta} - \beta}{\sqrt{Var(\hat{\beta})}} = \frac{\hat{\beta} - \beta}{\sigma\sqrt{\frac{1}{S_{xx}}}} \sim \mathcal{N}(0, 1).$$

Furthermore, $(\hat{\alpha}, \hat{\beta})$ and S^2 are independent and

$$\frac{(n-2)S^2}{\sigma^2} \sim \chi^2_{n-2},$$

$$\frac{\hat{\alpha} - \alpha}{S\sqrt{\frac{1}{n} + \frac{(\bar{x})^2}{S_{xx}}}} \sim t_{n-2},$$

$$\frac{\hat{\beta} - \beta}{S\sqrt{\frac{1}{S_{xx}}}} \sim t_{n-2}, \quad \text{or} \quad \frac{(\hat{\beta} - \beta)^2}{\frac{S^2}{S_{xx}}} \sim F_{1,n-2}.$$

Note that these results can be used to construct $100(1-\alpha)\%$ confidence intervals for α and β, which are given by

$$\alpha: \quad \Big(\hat{\alpha} - t_{n-2,\alpha/2}s\sqrt{\frac{1}{n} + \frac{\bar{x}^2}{S_{xx}}},\ \hat{\alpha} + t_{n-2,\alpha/2}s\sqrt{\frac{1}{n} + \frac{\bar{x}^2}{S_{xx}}}\Big);$$

$$\beta: \quad \Big(\hat{\beta} - t_{n-2,\alpha/2}s/\sqrt{S_{xx}},\ \hat{\beta} + t_{n-2,\alpha/2}s/\sqrt{S_{xx}}\Big),$$

where $t_{n-2,\alpha/2}$ satisfies that $\Pr(t_{n-2} \geq t_{n-2,\alpha/2}) = \alpha/2$.

Example: (Predicting height) Find $\hat{\sigma}^2$ and the CIs for $\hat{\beta}$ and $\hat{\alpha}$.

Solution: We have

$$\bar{x} = 34.286, \quad \bar{y} = 66.857, \quad n = 14,$$
$$\sum_{i=1}^{n} x_i^2 = 16558, \quad \sum_{i=1}^{n} y_i^2 = 62674, \quad \sum_{i=1}^{n} x_i y_i = 32183,$$
$$S_{xx} = 100.86, \; S_{xy} = 91.57, \; \hat{\beta} = 0.9079, \; \hat{\alpha} = 35.73.$$

Hence

$$\begin{aligned}
S_{yy} &= \sum_{i=1}^{n} y_i^2 - n(\bar{y})^2 = 62674 - 14(66.857^2) = 95.71, \\
s^2 &= \frac{S_{yy} - \hat{\beta}\, S_{xy}}{n-2} = \frac{95.71 - \; 0.9079(91.57)}{14-2} = 1.0478, \\
CI \text{ for } \beta &= \left(\hat{\beta} - t_{n-2,0.975}\sqrt{\frac{s^2}{S_{xx}}}, \; \hat{\beta} + t_{n-2,0.975}\sqrt{\frac{s^2}{S_{xx}}}\right) \\
&= \left(0.9079 - 2.1788\sqrt{\frac{1.0478}{100.86}}, \; 0.9079 + 2.1788\sqrt{\frac{1.0478}{100.86}}\right) \\
&= (0.6859, \; 1.13) \\
CI \text{ for } \alpha &= \left(\hat{\alpha} - t_{n-2,0.975}\sqrt{s^2\left(\frac{1}{n} + \frac{\bar{x}^2}{S_{xx}}\right)}, \; \hat{\alpha} + t_{n-2,0.975}\sqrt{s^2\left(\frac{1}{n} + \frac{\bar{x}^2}{S_{xx}}\right)}\right) \\
&= \left(35.73 - 2.1788\sqrt{1.0478\left(\frac{1}{14} + \frac{34.29^2}{100.86}\right)}, \right. \\
&\qquad \left. 35.73 + 2.1788\sqrt{1.0478\left(\frac{1}{14} + \frac{34.29^2}{100.86}\right)}\right) \\
&= (28.0906, \; 43.3655)
\end{aligned}$$

Since the CI for β does not contain 0, β is significantly greater than 0.

In R,

```
> Syy=sum(y^2)-sum(y)^2/n
> Syy
[1] 95.71429
> SSR=Syy-beta*Sxy
> SSR
[1] 12.57365
> s2=SSR/(n-2)
> s2
[1] 1.047805
> CIb.lower=beta-qt(0.975,n-2)*sqrt(s2/Sxx)
> CIb.upper=beta+qt(0.975,n-2)*sqrt(s2/Sxx)
> c(CIb.lower,CIb.upper)
[1] 0.6858534 1.130011
> CIa.lower=alpha-qt(0.975,n-2)*sqrt(s2*(1/n+mean(x)^2/Sxx))
> CIa.upper=alpha+qt(0.975,n-2)*sqrt(s2*(1/n+mean(x)^2/Sxx))
> c(CIa.lower,CIa.upper)
[1] 28.09063 43.36546
```

25.3 Test on β

Suppose that

$$Y_i = \alpha + \beta\, x_i + \epsilon_i, \qquad \epsilon_i \sim \mathcal{N}(0, \sigma^2).$$

If Y does not change with X, $\beta = 0$. Thus one may wish to test:

$$H_0: \ \beta = 0 \text{ vs } H_1: \ \beta \neq 0$$

in order to assess whether the explanatory variable X has an influence on the dependent variable Y.

The five steps to test the significance of β (i.e. $\beta \neq 0$) and hence the regression model are

1. **Hypotheses:** $H_0 : \beta = \beta_0$ vs $H_1 : \beta > \beta_0$, $\beta < \beta_0$, $\beta \neq \beta_0$.
2. **Test statistic:** $t_0 = \dfrac{\hat{\beta} - \beta_0}{s/\sqrt{S_{xx}}}$ where
$$s^2 = \frac{S_{yy} - \hat{\beta}\, S_{xy}}{n-2} \quad \text{and} \quad \hat{\beta} = S_{xy}/S_{xx}.$$
3. **Assumptions:** $Y_i \sim \mathcal{N}(\alpha + \beta x_i, \sigma^2)$. Y_i are independent.
4. ***P*-value:** $\Pr(t_{n-2} \geq t_0)$ for $H_1: \ \beta > \beta_0$,
$\Pr(t_{n-2} \leq t_0)$ for $H_1: \ \beta < \beta_0$;
$2\Pr(t_{n-2} \geq |t_0|)$ for $H_1 : \beta \neq \beta_0$.
5. **Decision:** Reject H_0 if p-value $< \alpha$.

Remarks

1. t_0 can be calculated by
$$t_0 = \frac{\sqrt{n-2}(\hat{\beta} - \beta_0)}{\sqrt{(n-2)s^2/S_{xx}}} = \frac{\sqrt{n-2}(S_{xy}/S_{xx} - \beta_0)}{\sqrt{(S_{xx}S_{yy} - S_{xy}^2)/S_{xx}^2}}.$$

In particular, if $\beta_0 = 0$,

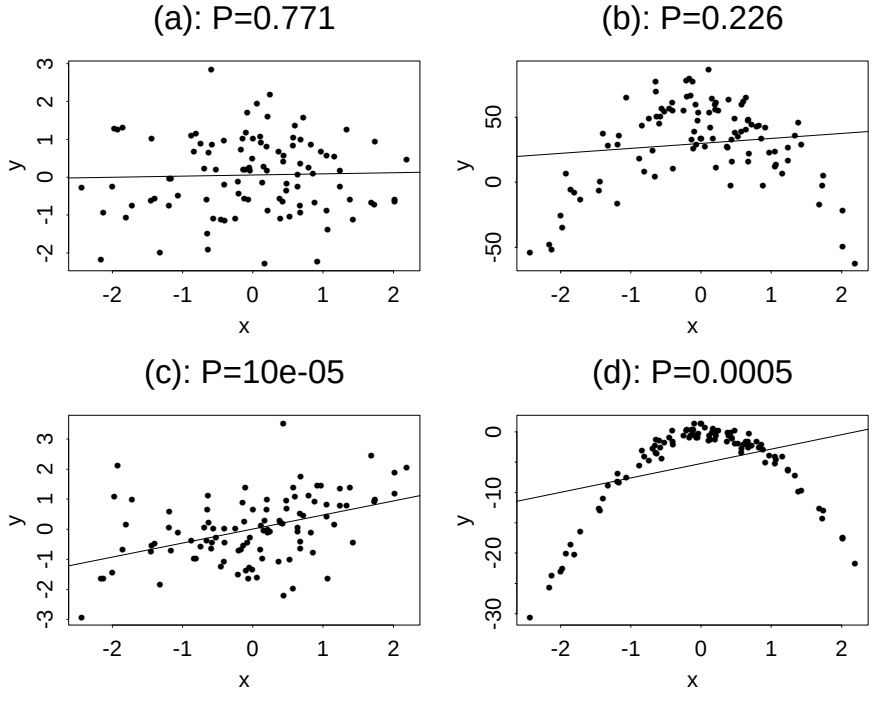
$$t_0 = \frac{\sqrt{n-2}\, S_{xy}}{\sqrt{S_{xx}S_{yy} - S_{xy}^2}} = \frac{\sqrt{n-2}\,\frac{S_{xy}}{\sqrt{S_{xx}S_{yy}}}}{\sqrt{1 - \frac{S_{xy}^2}{S_{xx}S_{yy}}}} = \frac{\sqrt{n-2}\, r}{\sqrt{1-r^2}}$$

where r is the *correlation coefficient* in the next section.

2. If $H_0 : \beta = 0$ is accepted, it implies that X has no linear effect in describing the behavior of the response Y. Note that it is not equivalent to say there are no relationship between X and Y. The true model may involve *quadratic*, *cubic*, or other more complex functions of X.

   The following examples illustrate the idea. The p-value indicated on the top of each plot corresponds to the 2-sided t-test for a zero slope.

   Case (b) shows insignificant result but there is a clear quadratic relationship between X and Y.



3. If $H_0 : \ \beta = 0$ is rejected, it implies that the data indicates there is a linear relationship between X and Y. However a linear regression might not be the best model to describe the relationship.

   Case (d) shows significant result but it is clear that a quadratic model is better to describe the relationship between X and Y.

4. The strength of linear relationship may be described by the sample correlation coefficient.

5. Similar ideas can be applied to construct a test for the intercept α. However there is more interest in testing β than α. The β contains the information about whether linear relationship exists between Y and X.

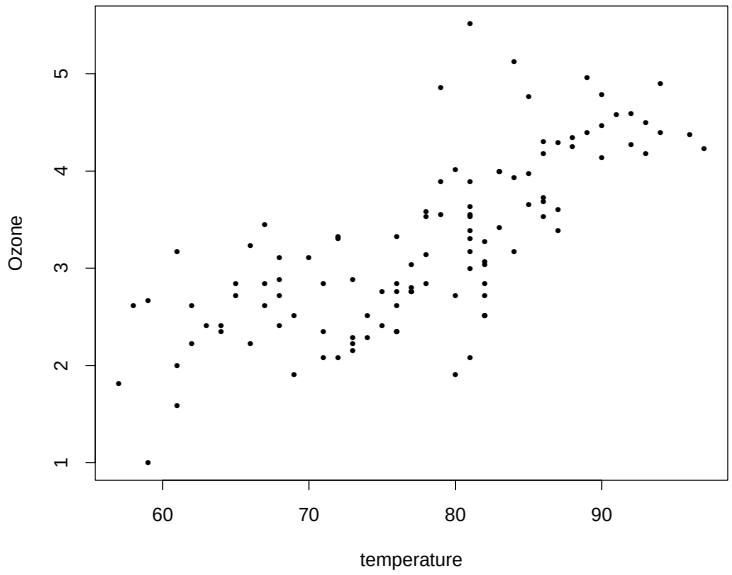
6. The regression line is sensitive to outliers as will be demonstrated in the computer practical. The nonparameteric version of the regression model is the *kernel smoothing* which is more robust to outliers but it is not included in this course.

Example: (Air pollution) The data `air` give four environmental variables Ozone, radiation, temperature and wind. We'd like to assess whether the variable temperature (X) has an influence on the dependent variable Ozone (Y).

```
>air
      ozone radiation temperature wind
 1 3.448217       190          67  7.4
 2 3.301927       118          72  8.0
 3 2.289428       149          74 12.6
 4 2.620741       313          62 11.5
...
110 2.620741       131          76  8.0
111 2.714418       223          68 11.5
```

Solution: First of all, we look at a scatter plot of the data to see whether a linear model is appropriate.

```
>plot(air[,3], air[,1])
```



The scatter plot indicates that a linear model is appropriate.

We fit a simple linear model

$$Y_i = \alpha + \beta x_i + \epsilon_i, \quad \epsilon_i \sim \mathcal{N}(0, \sigma^2)$$

to the data to assess whether the temperature (X) has an influence on Ozone (Y).

We have obtained the following values from R,

$$n = 111, \quad S_{xx} = 9990.23, \quad S_{yy} = 87.21, \quad S_{xy} = 702.95.$$

The test for the significance of β is

1. **Hypotheses:** $H_0: \ \beta = 0$ vs $H_1: \ \beta \neq 0$.
2. **Test statistic:** $\hat{\beta} = \dfrac{S_{xy}}{S_{xx}} = \dfrac{702.95}{9990.23} = 0.07036375$

$$s^2 = \frac{S_{yy} - \hat{\beta} S_{xy}}{n-2} = \frac{87.21 - 0.07036375(702.95)}{109} = 0.3463101$$

$$t_0 = \frac{\hat{\beta}}{\sqrt{\frac{s^2}{S_{xx}}}} = \frac{0.07036375}{\sqrt{\frac{0.3463101}{9990.23}}} = 11.95$$

$$\text{or } \ t_0 = \frac{\sqrt{n-2}\, S_{xy}}{\sqrt{S_{xx}S_{yy} - S_{xy}^2}} = \frac{\sqrt{109}(702.947)}{\sqrt{9990.234(87.209) - 702.947^2}} = 11.95$$

3. **Assumptions:** $Y_i \sim \mathcal{N}(\alpha + \beta x_i, \sigma^2)$. Y_i are independent.
4. ***P*-value:** p-value $= 2\Pr(t_{109} \geq 11.95) \approx 0$
5. **Decision:** There are very strong evidence in the data to indicate a linear relationship between temperature and ozone.

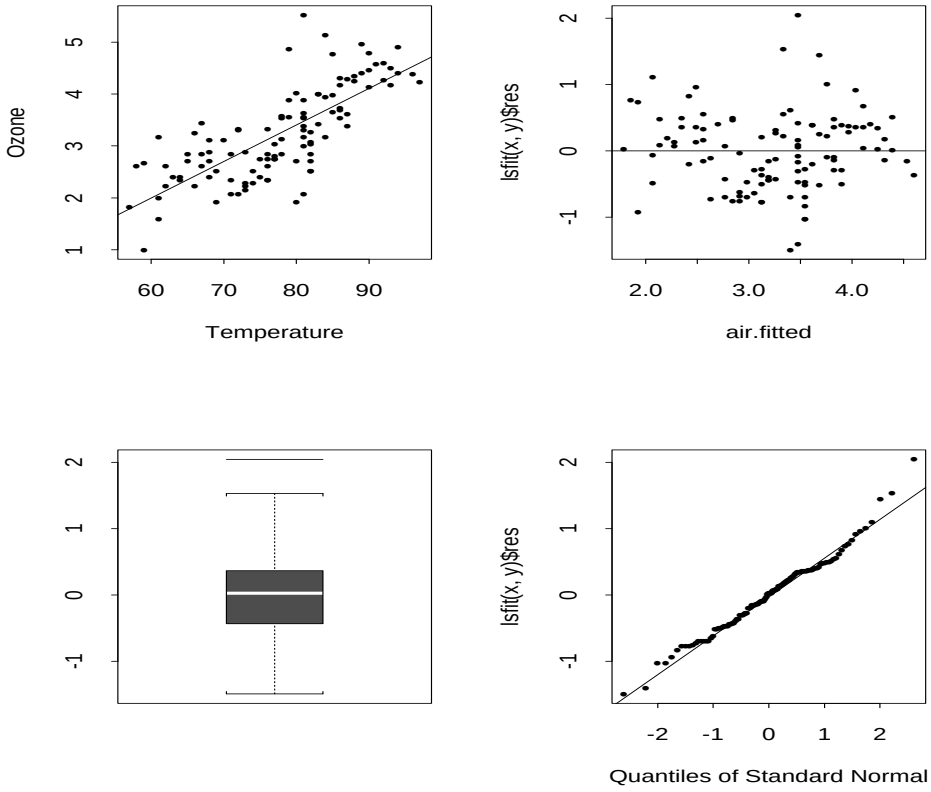
In R,

```
> x=air[,3]
> y=air[,1]

> c(mean(x),mean(y))
  77.792793  3.247784

> Sxx=sum((x-mean(x))*(x-mean(x)))
> Sxy=sum((x-mean(x))*(y-mean(y)))
> Syy=sum((y-mean(y))*(y-mean(y)))

> c(Sxx, Sxy, Syy)
  9990.23423  702.94721   87.20876

> n=length(x)
> beta=Sxy/Sxx  #check beta
> s2=(Syy-beta*Sxy)/(n-2)
> c(beta,s2)
[1] 0.07036344  0.3463025
> t0=beta/sqrt(s2/Sxx)
> t0=sqrt(n-2)*Sxy/sqrt(Sxx*Syy-Sxy^2)  #either way
> p=2*(1-pt(t0,n-2))
> p
[1] 0
```

The assumptions on the model may be checked as follows in R:

```
> lsfit(x,y)$coef
  Intercept          X
 -2.225984   0.07036344

> air.fitted=y- lsfit(x,y)$res  #fitted values
```

```
> par(mfrow=c(2,2))
> plot(x,y, xlab="Temperature",ylab="Ozone")  #fitted line plot
> abline(lsfit(x,y))
> plot(air.fitted, lsfit(x,y)$res)  #residual plot
> abline(h=0)
> boxplot(lsfit(x,y)$res)  #boxplot of residuals
> qqnorm(lsfit(x,y)$res)  #normal qq plot
> qqline(lsfit(x,y)$res)
```



Comments:

The scatter plot and the plot of residuals versus fitted values indicates that a linear model is appropriate. The boxplot and the normal qq-plot of residuals indicates that the distribution of residuals is slightly long tailed. There are three rather larger residuals in particular. This might indicate that a robust regression should be used.

26 Regression analysis: ANOVA and prediction

26.1 ANOVA for regression

The linear regression model

$$Y_i = \alpha + \beta\, x_i + \epsilon_i, \quad i = 1, 2, ..., n, \quad \text{where} \quad \epsilon_i \overset{\text{iid}}{\sim} \mathcal{N}(0, \sigma^2),$$

is *significant* if the slope parameter β that describes the effect of X on Y is non-zero.

Note that

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

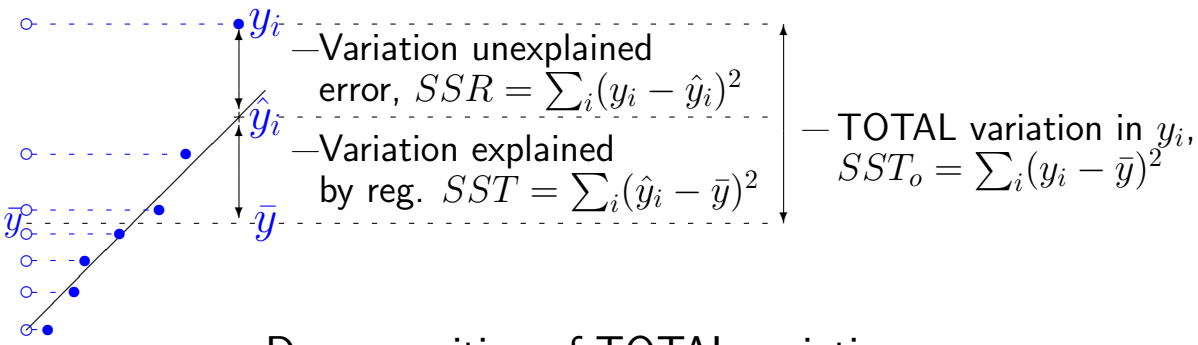
where $\hat{y}_i = \hat{\alpha} + \hat{\beta} x_i$. We have

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

where

$$\begin{aligned} SST &= \sum_{i=1}^{n}(\hat{y}_i - \bar{y})^2 = \sum_{i=1}^{n}(\hat{\alpha} + \hat{\beta} x_i - \bar{y})^2 = \sum_{i=1}^{n}(-\hat{\beta}\bar{x} + \hat{\beta} x_i)^2 \\ &= \hat{\beta}^2 \sum_{i=1}^{n}(x_i - \bar{x})^2 = \hat{\beta}\frac{S_{xy}}{S_{xx}} S_{xx} = \hat{\beta} S_{xy}, \\ SSR &= \sum_{i=1}^{n}(y_i - \hat{y}_i)^2 = (n-2)s^2 = S_{yy} - \hat{\beta}\, S_{xy}, \end{aligned}$$

since $\hat{\alpha} = \bar{y} - \hat{\beta}\bar{x} \Rightarrow \hat{\alpha} - \bar{y} = -\hat{\beta}\bar{x}$.



Decomposition of TOTAL variation

On the other hand, the F test statistic for testing $H_0 : \beta = \beta_0 = 0$ is

$$f_0 = t_0^2 = \frac{\hat{\beta}^2}{s^2/S_{xx}} = \frac{\hat{\beta}\frac{S_{xy}}{S_{xx}}S_{xx}}{s^2} = \frac{\hat{\beta}S_{xy}}{s^2} = \frac{SST/1}{SSR/(n-2)}.$$

Since $t^2_{m,1-\alpha/2} = F_{1,m,1-\alpha}$ for any degree of freedom m,

$$\begin{aligned} p\text{-value} &= \Pr\left(F_{1,n-2} \geq \frac{\hat{\beta}^2}{s^2/S_{xx}}\right) = \Pr\left(t^2_{n-2} \geq \frac{\hat{\beta}^2}{s^2/S_{xx}}\right) \\ &= \Pr\left(|t_{n-2}| \geq \frac{|\hat{\beta}|}{s/\sqrt{S_{xx}}}\right) = 2\Pr\left(t_{n-2} \geq \frac{|\hat{\beta}|}{s/\sqrt{S_{xx}}}\right). \end{aligned}$$

is the same as that in the t-test.

These calculations are summarized in the `Regression ANOVA table`, which is like the ANOVA tables obtained in the ANOVA test.

Regression ANOVA table

Source	df	SS	MS	F
Regression	1	S^2_{xy}/S_{xx}	MST	$\frac{MST}{s^2}$
Residuals	$n-2$	$S_{yy} - S^2_{xy}/S_{xx}$	$MSR = s^2$	
Total	$n-1$	S_{yy}		

Example: (Air pollution) Based on the previous calculations, we have

$$S_{xx} = 9990.234,\ S_{xy} = 702.947,\ S_{yy} = 87.209, \hat{\alpha} = -2.2260,\ \hat{\beta} = 0.07036$$

Complete the Regression ANOVA table and hence test for the significance of the regrssion model.

Solution: We have

$$\begin{aligned} SST &= \hat{\beta} S_{xy} = 0.07036(702.947) = 49.462 \\ SSR &= S_{yy} - SST = 87.209 - 49.462 = 37.747 \end{aligned}$$

Hence

Regression ANOVA table

Source	df	SS	MS	F
Regression	1	49.462	49.462	$\frac{49.462}{0.3463} = 142.828$
Residuals	109	37.747	$\frac{37.747}{109} = 0.3463$	
Total	110	87.209		

The test for the significance of the regression model is

1. **Hypotheses:** H_0: $\beta = 0$ vs H_1 : $\beta \neq 0$.
2. **Test statistic:** $f_0 = \frac{SST/1}{SSR/(n-2)} = 142.828 = 11.95^2 = t_0^2$.
3. **Assumption:** $Y_i \sim (\alpha + \beta x_i, \sigma^2)$. Y_i are independent.
4. **P-value:** p-value $= \Pr(F_{1,109} > 142.828) \approx 0$.
5. **Decision:** Since p-value < 0.05, we reject H_0. There are strong evidence of a linear relationship between the temperature (X) and Ozone (Y).

In R,

```
> summary(aov(air[,1]~air[,3]))

             Df Sum of Sq  Mean Sq  F Value Pr(F)
air[, 3]      1  49.46178 49.46178 142.8282     0
Residuals   109  37.74698  0.34630
```

Remarks

The test for the hypothesis that $\beta = 0$ in regression is similar to the ANOVA test for the completely randomized design (one-way data) that all treatments are equal. Under the null hypothesis, we have

$$\begin{aligned} \text{ANOVA test: } \mu_1 = \cdots = \mu_g = \mu &\Rightarrow Y_i \sim \mathcal{N}(\mu, \sigma^2), \\ \text{Reg. test: } \beta = 0 &\Rightarrow Y_i \sim \mathcal{N}(\alpha, \sigma^2), \end{aligned}$$

stating that the treatments (by groups in ANOVA or by explanatory variable X in regression) are unrelated to the response Y in any way.

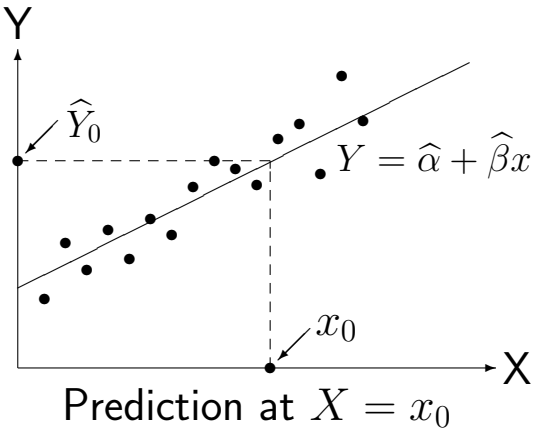
26.2 Estimation at $X = x_0$ (P.693-695)

Suppose we have fitted a simple linear regression model:

$$Y_i = \alpha + \beta\, x_i + \epsilon_i, \quad \text{where} \quad \epsilon_i \overset{\text{iid}}{\sim} \mathcal{N}(0, \sigma^2).$$

Obviously, $E(Y_i) = \alpha + \beta x_i$. It is often of interest to estimate the expected value $E(Y_0) = E(Y|X = x_0)$ at a specified value of $X = x_0$. An obvious choice for the estimate is

$$\hat{E}(Y_0) = \hat{\alpha} + \hat{\beta}\, x_0.$$



For this estimate, we have the following results (see P.229):

$$\begin{aligned}
E(\hat{\alpha} + \hat{\beta}x_0) &= \alpha + \beta\, x_0, \\
Var(\hat{\alpha} + \hat{\beta}x_0) &= Var(\hat{\alpha}) + x_0^2\, Var(\hat{\beta}) + 2x_0\, Cov(\hat{\alpha}, \hat{\beta}) \\
&= \sigma^2\Big(\frac{\sum_{i=1}^{n} x_i^2}{nS_{xx}} + \frac{x_0^2}{S_{xx}} - \frac{2x_0\bar{x}}{S_{xx}}\Big) \\
&= \sigma^2\Big[\frac{\Big(\sum_{i=1}^{n} x_i^2 - n\bar{x}^2\Big) + n\bar{x}^2}{nS_{xx}} + \frac{x_0^2}{S_{xx}} - \frac{2x_0\bar{x}}{S_{xx}}\Big] \\
&= \sigma^2\Big(\frac{S_{xx}}{nS_{xx}} + \frac{\bar{x}^2 - 2\bar{x}x_0 + x_0^2}{S_{xx}}\Big) \\
&= \sigma^2\left[\frac{1}{n} + \frac{(x_0 - \bar{x})^2}{S_{xx}}\right].
\end{aligned}$$

Hence

$$\frac{\widehat{E}(Y_0) - Y_0}{se[\widehat{E}(Y_0)]} = \frac{\hat{\alpha} + \hat{\beta}\, x_0 - (\alpha + \beta\, x_0)}{s\sqrt{\frac{1}{n} + \frac{(x_0 - \bar{x})^2}{S_{xx}}}} \sim t_{n-2},$$

and a $100(1-\alpha)\%$ confidence interval for $E(Y_0) = \alpha + \beta\, x_0$ is:

$$\left(\hat{\alpha} + \hat{\beta}\, x_0 - t_{n-2,1-\alpha/2} s \sqrt{\frac{1}{n} + \frac{(x_0 - \bar{x})^2}{S_{xx}}},\ \hat{\alpha} + \hat{\beta}\, x_0 + t_{n-2,1-\alpha/2} s \sqrt{\frac{1}{n} + \frac{(x_0 - \bar{x})^2}{S_{xx}}} \right)$$

where $t_{n-2,1-\alpha/2}$ satisfies that $\Pr(t_{n-2} \leq t_{n-2,1-\alpha/2}) = 1 - \alpha/2$.

Note: When $x_0 = 0$, $\hat{y}_0 = \hat{\alpha} + \hat{\beta} \times 0 = \hat{\alpha}$ and the estimation interval becomes the CI for α given by

$$\left(\hat{\alpha} - t_{n-2,1-\alpha/2} s \sqrt{\frac{1}{n} + \frac{\bar{x}^2}{S_{xx}}},\ \hat{\alpha} + t_{n-2,1-\alpha/2} s \sqrt{\frac{1}{n} + \frac{\bar{x}^2}{S_{xx}}} \right)$$

26.3 Prediction at $X = x_0$ (P.571-572)

Now instead of estimating the expected value $E(Y_0)$ at x_0, we wish to *predict* the response Y_0 that we will observe in the future when $X = x_0$.

Under the regression model of $Y_i = \alpha + \beta x_i + \epsilon_i$, we can employ

$$\hat{Y}_0 = \hat{\alpha} + \hat{\beta}\, x_0 + \hat{\epsilon}_0 = \hat{\alpha} + \hat{\beta}\, x_0,$$

as a predictor of $Y_0 = \alpha + \beta x_0 + \epsilon_0$ since $\hat{\epsilon}_0 = 0$. This is same as $\hat{E}(Y_0) = \hat{\alpha} + \hat{\beta}\, x_0$. Note that $\hat{Y}_0$ is unbiased since

$$E(\hat{Y}_0) = E(\hat{\alpha}) + E(\hat{\beta})\, x_0 + E(\hat{\epsilon}_0) = \alpha + \beta x_0 = E(Y_0)$$

since $E(\hat{\epsilon}_0) = 0$. However the variance of $\hat{Y}_0$ is greater than that of $\hat{E}(Y_0) = \hat{\alpha} + \hat{\beta}\, x_0$ because $\hat{Y}_0$ is predicted with prediction error ϵ_0:

$$\begin{aligned} Var(\hat{Y}_0) &= Var(\hat{\alpha} + \hat{\beta}\, x_0) + Var(\epsilon_0) \\ &= \sigma^2\Big[\frac{1}{n} + \frac{(x_0 - \bar{x})^2}{S_{xx}}\Big] + \sigma^2 \\ &= \sigma^2\Big[1 + \frac{1}{n} + \frac{(x_0 - \bar{x})^2}{S_{xx}}\Big] \end{aligned}$$

and $\hat{\alpha} + \hat{\beta}\, x_0$ and ϵ_0 are independent. Hence

$$\frac{\widehat{Y}_0 - Y_0}{se(\widehat{Y}_0)} = \frac{\hat{\alpha} + \hat{\beta}\, x_0 - (\alpha + \beta x_0)}{s\sqrt{1 + \frac{1}{n} + \frac{(x_0 - \bar{x})^2}{S_{xx}}}} \sim t_{n-2},$$

and a $100(1-\alpha)\%$ prediction intervals for Y_0 is given by

$$\left(\hat{\alpha} + \hat{\beta}\, x_0 - t_{n-2,1-\frac{\alpha}{2}} s\sqrt{1 + \frac{1}{n} + \frac{(x_0 - \bar{x})^2}{S_{xx}}},\ \hat{\alpha} + \hat{\beta}\, x_0 + t_{n-2,1-\frac{\alpha}{2}} s\sqrt{1 + \frac{1}{n} + \frac{(x_0 - \bar{x})^2}{S_{xx}}}\right)$$

where $t_{n-2,1-\frac{\alpha}{2}}$ satisfies that $\Pr(t_{n-2} \le t_{n-2,1-\frac{\alpha}{2}}) = 1 - \frac{\alpha}{2}$.

26.4 Effects on confidence intervals

Note: $$\text{Var}(\widehat{\alpha}) = \frac{\sigma^2 \sum_{i=1}^{n} x_i^2}{nS_{xx}}, \quad \text{Var}(\widehat{\beta}) = \frac{\sigma^2}{S_{xx}},$$

$$\text{Var}[E(\widehat{Y}_0)] = \sigma^2 \left(\frac{1}{n} + \frac{(x_0 - \bar{x})^2}{S_{xx}} \right) \text{ and } \text{Var}(\widehat{Y}_0) = \sigma^2 \left(1 + \frac{1}{n} + \frac{(x_0 - \bar{x})^2}{S_{xx}} \right)$$

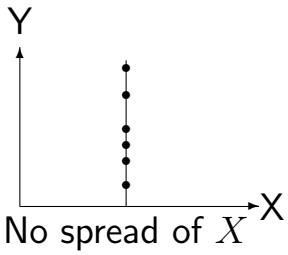
all depend on summary statistics

$$\sum_{i=1}^{n} Y_i, \quad \sum_{i=1}^{n} x_i, \quad \sum_{i=1}^{n} Y_i^2, \quad \sum_{i=1}^{n} x_i Y_i, \quad \sum_{i=1}^{n} x_i^2$$

which vary from sample to sample. Hence $\widehat{\alpha}$, $\widehat{\beta}$, $E(\widehat{Y}_0)$ and $\widehat{Y}_0$ are random variables which vary from sample to sample. Their variances depend on the following factors:

1. The smaller the σ^2 (with estimate S^2 which measures the variability of residuals ϵ_i), the better the fit and hence the smaller the variances for $\widehat{\alpha}$, $\widehat{\beta}$, $E(\widehat{Y}_0)$ and $\widehat{Y}_0$.

2. The larger the spread of x_i as measured by S_{xx}, the more information about Y_i and hence the smaller the variances for $\widehat{\alpha}$, $\widehat{\beta}$, $E(\widehat{Y}_0)$ and $\widehat{Y}_0$.

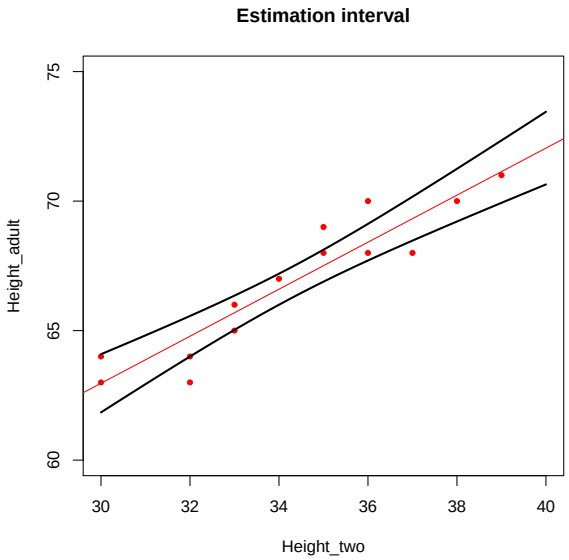


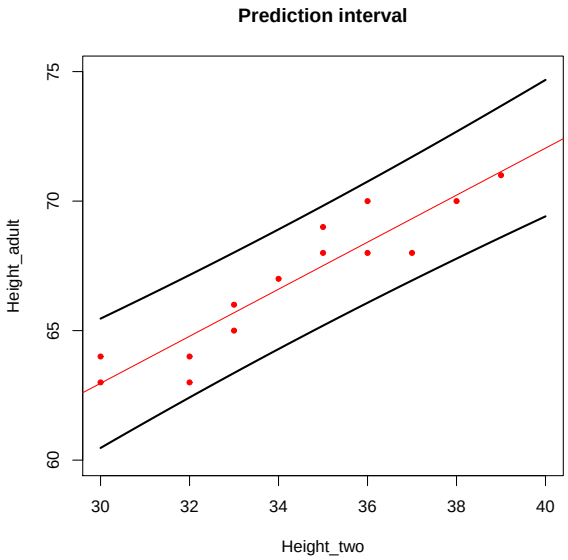
If all x are the same, it would be not be able to draw any conclusion about the dependence of Y on x.

In this case, $S_{xx} = 0$ giving $Var(\beta) = \infty$.

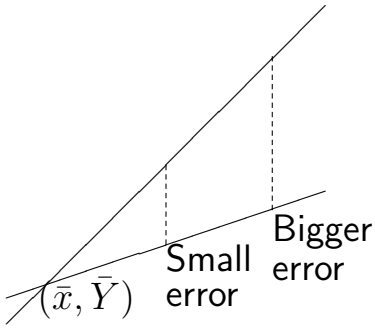
3. The larger the sample size n giving more information about Y_i, the smaller the variances for $\widehat{\alpha}$, $E(\widehat{Y}_0)$ and $\widehat{Y}_0$.

4. The closer is x_0 from $\bar{x}$, the smaller the squared distance $(x_0 - \bar{x})^2$, the smaller the variances for $E(\widehat{Y}_0)$ and $\widehat{Y}_0$.





The interval is shorter in the middle and the estimation interval is shorter than the prediction interval.



Why does the error increase away from $\bar{x}$?
Because if you wiggle the regression line, it makes more of a difference as it gets further away from the mean $\bar{x}$. Note that the line always passes through $(\bar{x}, \bar{Y})$.

Since when $X = \bar{x}$,

$$\begin{aligned} Y &= \widehat{\alpha} + \widehat{\beta}\bar{x} \\ &= (\bar{Y} - \widehat{\beta}\bar{x}) + \widehat{\beta}\bar{x} \\ &= \bar{Y} \end{aligned}$$

Hence $(\bar{x}, \bar{Y})$ passes through the line $Y = \widehat{\alpha} + \widehat{\beta}\bar{x}$.

Example: (Predicting height) Parents are often interested in predicting the eventual heights of their children. The following is a portion of the data taken from a study of heights of boys.

Height (inches)	39	30	32	34	35	36	36	30
at age two (x_i)	33	37	33	38	32	35		
Height (inches)	71	63	63	67	68	68	70	64
as an adult (y_i)	65	68	66	70	64	69		

1. Indicate whether a linear relationship between the heights of boys at age two and as an adult is significant.
2. Estimate the *average height of adults* that were 40 inches tall at the age of two and give a 90% *estimation interval* of this estimate.
3. Predict the *height of an adult* who was 40 inches tall at the age of two. Give a 90% *prediction interval* of this prediction.

Solution: From previous working, we have obtained that

$$\bar{x} = 34.286, \quad S_{xx} = 100.86, \quad S_{yy} = 95.71, \quad S_{xy} = 91.57,$$

and the least squares line is

$$\hat{y} = \hat{\alpha} + \hat{\beta}\, x = 35.73 + 0.9079x.$$

1. Then

$$\begin{aligned} SST &= S_{xy}^2/S_{xx} = \frac{91.57^2}{100.86} = 83.14 \\ SSR &= S_{yy} - S_{xy}^2/S_{xx} = 95.71 - 83.14 = 12.57 \end{aligned}$$

Hence the regression ANOVA table is obtained as follow:

Source	df	SS	MS	F
Regression	1	83.14	83.14	$\frac{83.14}{1.0478} = 79.35$
Residuals	12	12.57	$\frac{12.57}{12} = 1.0478$	
Total	13	95.71		

The test for the significance of the regression model is

1. **Hypotheses:** H_0: $\beta = 0$ vs H_1 : $\beta \neq 0$.
2. **Test statistic:** $f_0 = 79.35$.
3. **Assumption:** $Y_i \sim (\alpha + \beta x_i, \sigma^2)$. Y_i are independent.
4. **P-value:** $\Pr(F_{1,12} > 79.37) \approx 0$ ($F_{1,12,0.999} = 18.6$, 0.000 from R).
5. **Decision:** Since p-value < 0.05, we reject H_0. There are strong evidence of a linear relationship between the heights of boys at age two and the heights as an adult.

In R,

```
> Syy=sum(y^2)-sum(y)^2/n
> Syy
[1] 95.71429
> SST=beta*Sxy
> SST
[1] 83.14063
> SSR=Syy-SST
> SSR
[1] 12.57365
> s2=SSR/(n-2)
> s2
[1] 1.047805
> f0=SST/s2
> f0
```

```
[1] 79.34746
> t0=sqrt(f0)
> t0
[1] 8.90772
> p.value=1-pf(f0,1,n-2)
> p.value
[1] 1.230876e-06
```

2. According to the fitted line, the expected average height $E(Y_0)$ as an adult corresponding to the height of 40 inches at age two is

$$\hat{E}(Y_0) = 35.73 + 0.9079 \times 40 = 72.045.$$

Note that $n = 14$, $t_{12,0.05} = 1.782$, $s^2 = 1.0478$, $s = 1.0236$. Hence, the 90% *estimation interval* of the average height of adults when their heights at age two are $x_0 = 40$ is

$$\left(\hat{\alpha}+\hat{\beta}x_0 - t_{n-2,\frac{\alpha}{2}} s\sqrt{\frac{1}{n} + \frac{(x_0-\bar{x})^2}{S_{xx}}},\ \hat{\alpha}+\hat{\beta}x_0 + t_{n-2,\frac{\alpha}{2}} s\sqrt{\frac{1}{n} + \frac{(x_0-\bar{x})^2}{S_{xx}}} \right)$$

$$= \Big(72.045 - 1.782(1.0236)\sqrt{\frac{1}{14} + \frac{(40-34.286)^2}{100.86}},$$

$$72.045 + 1.782(1.0236)\sqrt{\frac{1}{14} + \frac{(40-34.286)^2}{100.86}}\Big)$$

$$= (70.90,\ \ 73.19).$$

3. According to the fitted line, the predicted height of a boy with height 40 inches at age two is still 72.046 inches. The 90% *prediction interval* of the height of an adult when his height at age two is $x_0 = 40$ is

$$\left(\hat{\alpha}+\hat{\beta}x_0 - t_{n-2,\frac{\alpha}{2}}s\sqrt{1+\frac{1}{n}+\frac{(x_0-\bar{x})^2}{S_{xx}}},\right.$$
$$\left.\hat{\alpha}+\hat{\beta}x_0 + t_{n-2,\frac{\alpha}{2}}s\sqrt{1+\frac{1}{n}+\frac{(x_0-\bar{x})^2}{S_{xx}}}\right)$$

$$= \left(72.045 - 1.782(1.0236)\sqrt{1+\frac{1}{14}+\frac{(40-34.286)^2}{100.86}},\right.$$
$$\left.72.045 + 1.782(1.0236)\sqrt{1+\frac{1}{14}+\frac{(40-34.286)^2}{100.86}}\right)$$

$$= (69.89,\ 74.20).$$

In R,

```
> x0=40
> y0=alpha+beta*x0
> y0
[1] 72.04533
> alp=0.1
> t=qt(1-alp/2,n-2)
> t
[1] 1.782288
> se.est=sqrt(s2*(1/n+(x0-mean(x))^2/Sxx))
> se.est
[1] 0.6434872
```

```
> se.pre=sqrt(s2*(1+1/n+(x0-mean(x))^2/Sxx))
> se.pre
[1] 1.209082
> CIest.low=y0-t*se.est
> CIest.up=y0+t*se.est
> c(CIest.low,CIest.up)
[1] 70.89845 73.19220
> CIpre.low=y0-t*se.pre
> CIpre.up=y0+t*se.pre
> c(CIpre.low,CIpre.up)
[1] 69.89039 74.20026
```

Note: $x_0 = 40$ is outside the range of X in the data. If x_0 is well outside the range of X, the estimation and prediction may not be reliable because we are not sure if the fitted relationship can be extended to x_0.

27 Regression and correlation

27.1 Correlation coefficient (P.575-582,587-589)

In practice, one may be interested in strength of the linear relationship between X and Y. Suppose X and Y are two random variables. The *correlation coefficient* ρ, as a measure of the *strength* of the linear relationship between X and Y, is defined as:

$$\begin{aligned}\rho &= \frac{E\{[X-E(X)][(Y-E(Y)]\}}{\sqrt{Var(x)Var(Y)}} = \frac{E[XY-E(X)E(Y)]}{\sqrt{Var(x)Var(Y)}} \\ &= \frac{Cov(X,Y)}{\sqrt{Var(x)Var(Y)}}.\end{aligned}$$

Note: If $\rho = 0$, then X and Y are said to be *uncorrelated*.

Definition: The *sample correlation coefficient*, r, for the bivariate data

$$(x_1, y_1),\ (x_2, y_2),\ \ldots,\ (x_n, y_n),$$

is defined as

$$r = \frac{S_{xy}}{\sqrt{S_{xx}\, S_{yy}}} = \frac{\sum_{i=1}^{n}(x_i-\bar{x})(y_i-\bar{y})}{\sqrt{\left[\sum_{i=1}^{n}(x_i-\bar{x})^2\right]\left[\sum_{i=1}^{n}(y_i-\bar{y})^2\right]}} \text{ or}$$

$$r = \frac{\sum_{i=1}^{n} x_i y_i - n\bar{x}\bar{y}}{\sqrt{\left[\sum_{i=1}^{n} x_i^2 - n\bar{x}^2\right]\left[\sum_{i=1}^{n} y_i^2 - n\bar{y}^2\right]}}$$

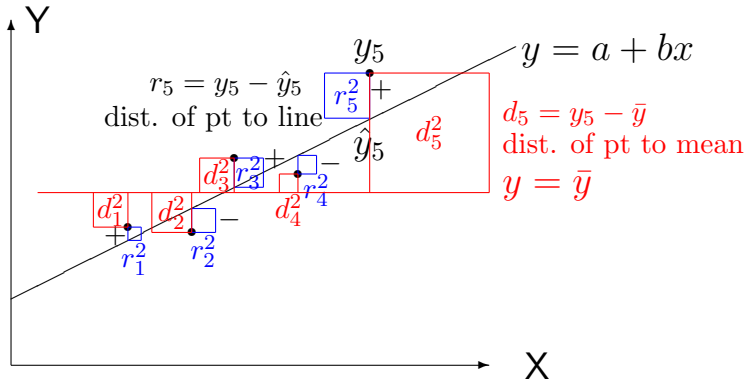
Hence r gives a point estimator of ρ, and hence measures the strength of the linear relationship between X and Y based on the sample.

27.2 Interpreting r and r^2

Given a bivariate data

$$(x_1, y_1),\ (x_2, y_2),\ \ldots,\ (x_n, y_n),$$

The square of correlation coefficient r^2 called the *coefficient of determination* measures the proportion of *total* variation in Y *explained* by the linear regression model:



Residuals

It is 'one minus the proportion of variation not explained by the model':

$$r^2 = 1 - \frac{\sum_i r_i^2}{\sum_i d_i^2} = 1 - \frac{\sum_i (y_i - \hat{y}_i)^2}{\sum_i (y_i - \bar{y})^2} = 1 - \frac{SSR}{SST_o}$$

where

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

$$\text{or} \quad \underbrace{S_{yy}}_{SST_o} = \underbrace{S_{xy}^2/S_{xx}}_{SST} + \underbrace{S_{yy} - S_{xy}^2/S_{xx}}_{SSR},$$

and SST_o is the *total* variation in Y,

SST is the sum of squares *explained* by the regression line and $SSR = SST_o - SST$ is the variation in Y remain *unexplained*.

Note that

$$r = \frac{S_{xy}}{\sqrt{S_{xx}\,S_{yy}}},$$

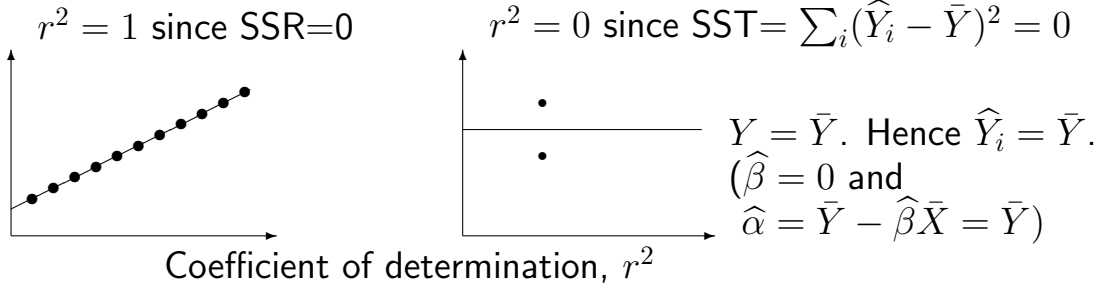
$$r^2 = \frac{S_{xy}^2}{S_{xx}\,S_{yy}} = \frac{S_{xy}^2/S_{xx}}{S_{yy}} = \frac{SST}{SST_o}, \quad \text{and}$$

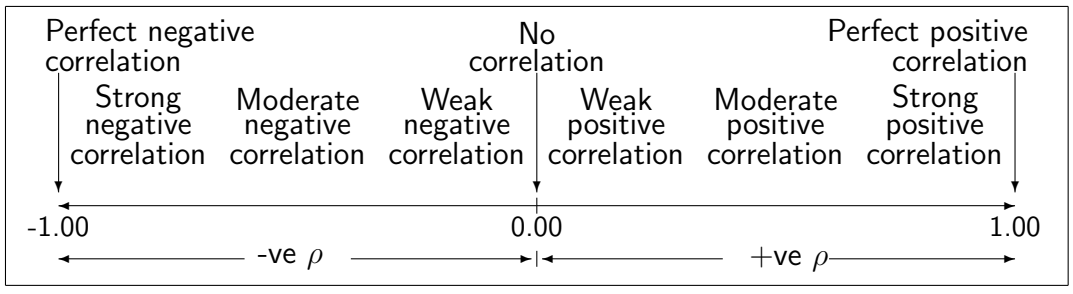
$$1 - r^2 = \frac{S_{yy} - S_{xy}^2/S_{xx}}{S_{yy}} = \frac{SSR}{SST_o}.$$

Hence the *coefficient of determination* r^2 measures the strength of the linear relationship between X and Y by the percentage of variation in Y explained by the linear regression model in X.

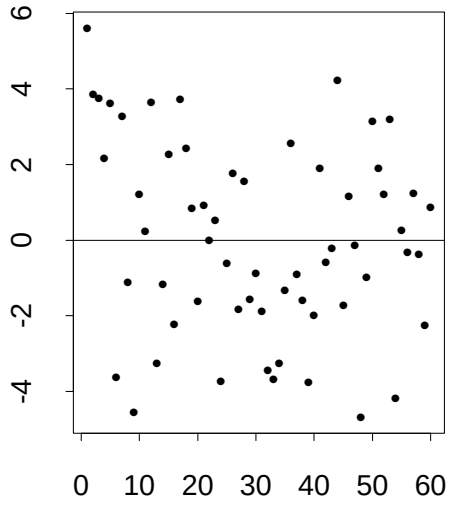
Interpreting r and r^2:

1. $|r| \leq 1$.
2. $r^2 = 1$ when all data in the scatterplot lie on the fitted least squares line.
3. $r = 0$ when there is no linear relationship between the points in the scatterplot.

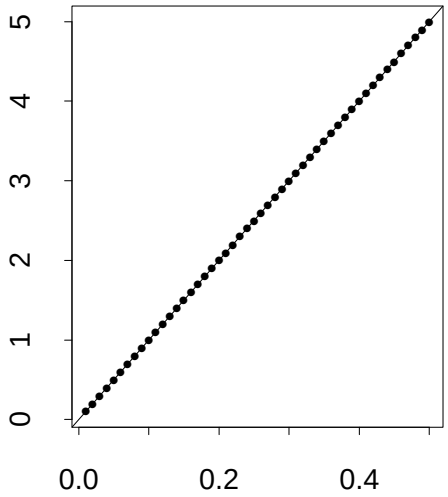




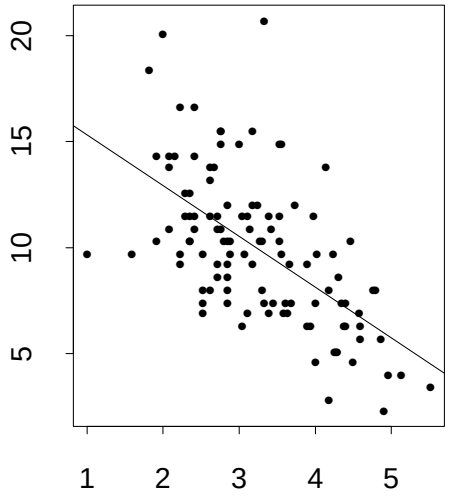
r=0



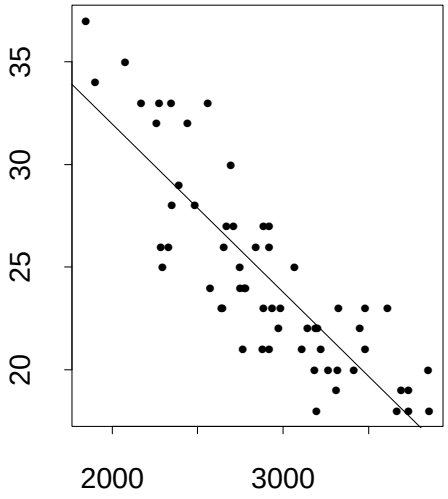
r=1



r=-0.6



r=-0.86



27.3 Test for ρ

By using the sample correlation coefficient r, we may construct a test for $\rho = 0$. The five steps for the test of the correlation coefficient ρ are:

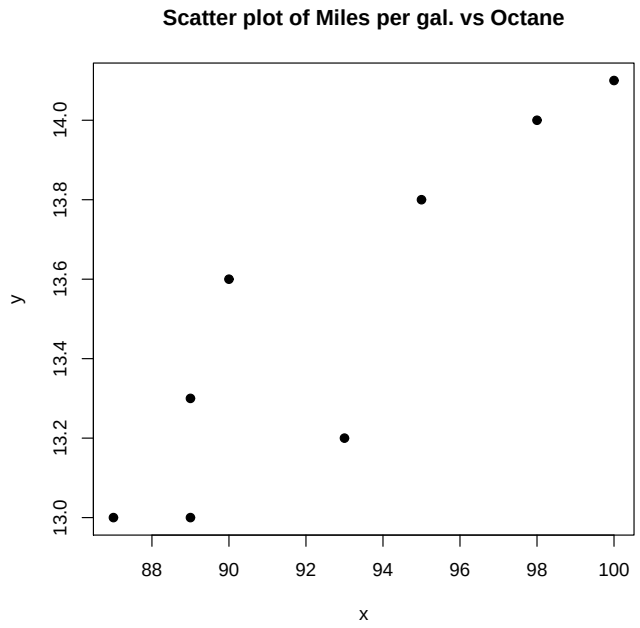
1. **Hypotheses:** $H_0 : \rho = 0$ vs $H_1 : \rho > 0$, $\rho < 0$, $\rho \neq 0$.
2. **Test statistic:** $t_0 = \dfrac{r\sqrt{n-2}}{\sqrt{1-r^2}}$.
3. **Assumption:** The data are taken from a bivariate normal population.
4. ***P*-value:** $\Pr(t_{n-2} \geq t_0)$ for $H_1 : \rho > 0$,
   $\Pr(t_{n-2} \leq t_0)$ for $H_1 : \rho < 0$,
   $2\Pr(t_{n-2} \geq |t_0|)$ for $H_1 : \rho \neq 0$.
5. **Decision:** Reject H_0 if p-value $< \alpha$.

Note: The tests for $\rho = 0$ and for $\beta = 0$ are equivalent since

$$t_0^2 = \frac{r^2(n-2)}{1-r^2} = \frac{\frac{S_{xy}^2}{S_{xx}S_{yy}}(n-2)}{1-\frac{S_{xy}^2}{S_{xx}S_{yy}}} = \frac{\frac{S_{xy}^2}{S_{xx}}(n-2)}{S_{yy}-\frac{S_{xy}^2}{S_{xx}}} = \frac{\hat{\beta}S_{xy}(n-2)}{S_{yy}-\hat{\beta}S_{xy}} = \frac{SST}{SSR/(n-2)} = f_0.$$

Example: (Automobile) The data in the following table give the miles per gallon obtained by a test automobile when using gasolines of varying octane levels.

Miles per Gallon (y)	13	13.2	13	13.6	13.3	13.8	14.1	14
Octane (x)	89	93	87	90	89	95	100	98



Given

$$\sum_{i=1}^{8} x_i = 741,\ \sum_{i=1}^{8} y_i = 108,\ \sum_{i=1}^{8} x_i y_i = 10016.3,\ \sum_{i=1}^{8} x_i^2 = 68789,\ \sum_{i=1}^{8} y_i^2 = 1459.34$$

(a) Calculate the value r.

(b) Do the data provide sufficient evidence to indicate that the octane level and miles per gallon are dependent?

Solution: We have that

$$S_{xx} = \sum_{i=1}^{n} x_i^2 - \frac{1}{n}(\sum_{i=1}^{n} x_i)^2 = 68789 - \frac{1}{8}741^2 = 153.875,$$

$$S_{yy} = \sum_{i=1}^{n} y_i^2 - \frac{1}{n}(\sum_{i=1}^{n} y_i)^2 = 1459.34 - \frac{1}{8}108^2 = 1.34,$$

$$S_{xy} = \sum_{i=1}^{n} x_i y_i - \frac{1}{n}(\sum_{i=1}^{n} x_i)(\sum_{i=1}^{n} y_i) = 10016.3 - \frac{1}{8}(741)(108) = 12.8.$$

(a) $r = S_{xy}/\sqrt{S_{xx}S_{yy}} = 12.8/\sqrt{153.875(1.34)} = 0.8914019.$

(b) The test for the correlation coefficient ρ between the octane level and miles per gallon is

1. **Hypothesis:** $H_0 : \rho = 0 \quad \text{vs} \quad H_1 : \rho \neq 0.$
2. **Test statistic:** $t_0 = \dfrac{r\sqrt{n-2}}{\sqrt{1-r^2}} = \dfrac{0.8914019\sqrt{8-2}}{\sqrt{1-0.8914019^2}} = 4.81777.$
3. **Assumption:** The data are taken from a bivariate normal population.
4. ***P*-value:** $p\text{-value} = 2\ \Pr(t_6 \geq 4.82) \in (0.002,\ 0.01)$
   $(t_{6,0.995} = 3.707,\ t_{6,0.999} = 5.208;\ 0.00294677 \text{ from R})$
5. **Decision:** Since the p-value < 0.05, we reject H_0. There are strong evidence in the data that the octane level and miles per gallon are dependent.

In R,

```
> x=c(89, 93, 87, 90, 89, 95, 100, 98)
> y=c(13, 13.2, 13, 13.6, 13.3, 13.8, 14.1, 14)
> cor.test(x,y,alt="two.sided")
```

```
        Pearson's product-moment correlation

data:  x and y
t = 4.8178, df = 6, p-value = 0.002947
alternative hypothesis: true correlation is not equal to 0
95 percent confidence interval:
 0.5021576 0.9803015
sample estimates:
      cor
0.8914019

> n=length(x)   #checking
> Sxx=sum(x^2)-sum(x)^2/n
> Sxy=sum(x*y)-sum(x)*sum(y)/n
> Syy=sum(y^2)-sum(y)^2/n
> c(Sxx,Sxy,Syy)
[1] 153.875  12.800   1.340
> beta=Sxy/Sxx
> alpha=mean(y)-beta*mean(x)
> c(alpha,beta)
[1] 5.7950447 0.0831844
> SST=beta*Sxy
> r2=SST/Syy
> r=sqrt(r2)
> t0=r*sqrt((n-2)/(1-r^2))
> p.value=2*(1-pt(abs(t0),n-2))
> c(SST,r2,r,t0,p.value)
[1] 1.064760357 0.794597282 0.891401863 4.817770070 0.00294677
```