MATH 204: Principles of Statistics 2

Winter 2026

Dr David A. Stephens

Department of Mathematics & Statistics
Room 1225, Burnside Hall

david.stephens@mcgill.ca
www.math.mcgill.ca/dstephens/204/

Textbook: McClave and Sincich, Statistics (13th Edition) Chapters 10-14.

Note: Recent editions of McClave and Sincich contain essentially the same material.

Prerequisites: MATH 203 (or equivalent)

Some statistical computing knowledge useful.

Method of Assessment:

- Assignments
- Quizzes
- Mid-Term
- Final

See syllabus handout for precise details.

Course Objectives

- Extension of MATH 203 topics to other practical experimental contexts
- Basic statistical computation using software (R)
- Practice in the use of statistical methods, in particular, hypothesis testing and linear modelling.

Course Objectives

From the Course Catalogue:

The concept of degrees of freedom and the analysis of variability. Planning of experiments. Experimental designs. Polynomial and multiple regressions. Statistical computer packages (no previous computing experience is needed). General statistical procedures requiring few assumptions about the probability model.

Three main parts

I. The Analysis of Variance and Designed Experiments
   - *The concept of degrees of freedom*
   - *The analysis of variability*
   - *Planning of experiments*
   - *Experimental designs*

II. Linear Regression Modelling
   - *Polynomial and multiple regressions*

III. Non-Parametric Testing
   - *General statistical procedures requiring few assumptions about the probability model*

Statistical computer packages (no previous computing experience is needed)

Recap of hypothesis testing

Typical experimental scenario:

- two different groups of subjects
- single observation/measurement made on each subject
- scientific question of interest

Example: Pre-Natal Care

Comparison of the birthweights of babies in two groups of mothers:

- GROUP A: Received five or fewer pre-natal visits
- GROUP B: Received more than five pre-natal visits

Do the GROUP A babies have

significantly different

birthweights from those from GROUP B ?

Data: Birthweights (grammes)

- GROUP A: 10 subjects

2164	2600	2184	2080	1820
2496	2184	2080	2184	2576

- GROUP B: 7 subjects

3224	2704	2912	2444	3120
		2912	3848	

First step in analysis: statistical summary

- GROUP A:
  - Sample size: $n_A = 10$
  - Sample mean: $\overline{x}_A = 2236.8$
  - Sample variance: $s_A^2 = 61190.4$
- GROUP B:
  - Sample size: $n_B = 7$
  - Sample mean: $\overline{x}_B = 3023.429$
  - Sample variance: $s_B^2 = 198679.6$

First step in analysis: statistical summary

Recall, for data $x_1, \ldots, x_n$

$$\begin{aligned} \overline{x} &= \frac{x_1 + \cdots + x_n}{n} = \frac{1}{n}\sum_{i=1}^{n} x_i \\ s^2 &= \frac{1}{n-1}\sum_{i=1}^{n}(x_i - \overline{x})^2 \end{aligned}$$

$\overline{x}$ measures the average value of sample

s^2 measures the amount of variability around the average.

First step in analysis: statistical summary

Implementation in R:

```
> load("204DataSets.rda")
> Birthweight

##    Group gp   BW
## 1      A  1 2164
## 2      A  1 2600
## 3      A  1 2184
## 4      A  1 2080
## 5      A  1 1820
## 6      A  1 2496
## 7      A  1 2184
## 8      A  1 2080
## 9      A  1 2184
## 10     A  1 2576
## 11     B  2 3224
## 12     B  2 2704
## 13     B  2 2912
## 14     B  2 2444
## 15     B  2 3120
## 16     B  2 2912
## 17     B  2 3848
```

First step in analysis: statistical summary

```
> xA<-Birthweight$BW[Birthweight$Group=='A']
> c(length(xA),mean(xA),var(xA))

## [1]    10.0  2236.8 61190.4

> xB<-Birthweight$BW[Birthweight$Group=='B']
> c(length(xB),mean(xB),var(xB))

## [1]      7.000   3023.429 198679.619
```

First step in analysis: statistical summary

In the birthweight example

$$\overline{x}_A = 2236.8 \qquad \overline{x}_B = 3023.429$$

so it appears that Group B birthweights are higher....

... but are they significantly higher ?

i.e. is the difference due to chance alone ? Clearly

- sample sizes quite small
- birthweights quite variable

Statistical testing

We adopt the following procedure to assess the significance of the difference between $\overline{x}_A$ and $\overline{x}_B$.

1. Define a test statistic, T, that permits comparison of the two groups.

2. Predict how T will behave assuming that the two groups are not significantly different.

3. Compare the prediction with what was actually observed.

Statistical testing

Formally, we

- assume a Normal distribution for the data

  Assume $x_{A1}, \ldots, x_{An_A}$ are drawn from a population that is well-modelled by a

$$Normal(\mu_A, \sigma_A^2)$$

  distribution. Similarly

$$x_{B1}, \ldots, x_{Bn_B} \sim Normal(\mu_B, \sigma_B^2)$$

  We might initially assume that

$$\sigma_A^2 = \sigma_B^2.$$

Statistical testing

- consider the two hypotheses

$$H_0 : \quad \mu_A = \mu_B$$
$$H_a : \quad \mu_A \neq \mu_B$$

H_0 is the null hypothesis
H_a is the alternative hypothesis

  - here H_a is a two-sided alternative
  - this leads to a two-sided test

Statistical testing

- find the distribution of T assuming that H_0 is true
  - this distribution is known as the null distribution.
  - for probability calculations associated with the null distribution, we use the notation P_0: for example, we might compute

$$P_0(T > 2.04) \qquad P_0(T \leqslant -1.13)$$

  and so on.

Statistical testing

- compute the observed test statistic

$$T_{obs} = \frac{\overline{x}_A - \overline{x}_B}{s_P \sqrt{\frac{1}{n_A} + \frac{1}{n_B}}}$$

where

$$s_P^2 = \frac{(n_A - 1)s_A^2 + (n_B - 1)s_B^2}{n_A + n_B - 2}$$

Statistical testing

s_P^2 is an estimate of the common population variance

$$\sigma^2 = \sigma_A^2 = \sigma_B^2.$$

Here

$$s_P^2 = \frac{(10-1)61190.4 + (7-1)198679.6}{10+7-2} = 116186.1$$

so that

$$s_P = 340.8608.$$

Thus

$$T_{obs} = \frac{2236.8 - 3023.429}{340.8608\sqrt{\frac{1}{10} + \frac{1}{7}}} = -4.683$$

Statistical testing

If the null hypothesis was true, so that $\mu_A = \mu_B$, the computed test statistic T_{obs} should look like an observation from a

Student-t or t

distribution with

$$n_A + n_B - 2 = 15$$

degrees of freedom. We write t_{15} to denote this distribution.

If H_0 is true, T_{obs} should lie somewhere in a

high probability region

of the t_{15} probability distribution.

Statistical testing

Calculation in R:

```
> nA<-length(xA)
> nB<-length(xB)
> sP.sq<-((nA-1)*var(xA)+(nB-1)*var(xB))/(nA+nB-2)
> sP.sq

+ [1] 116186.1

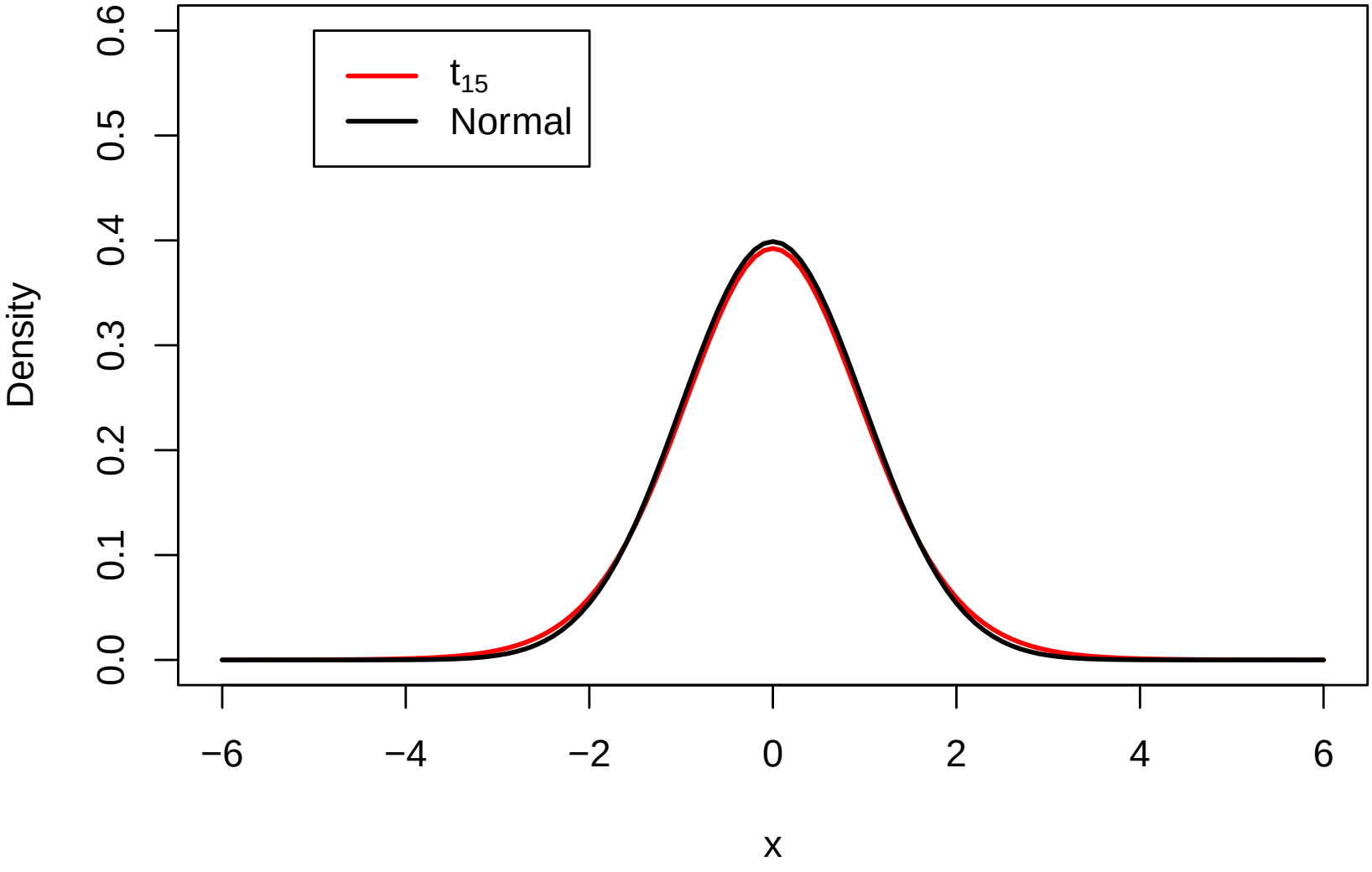
> sP<-sqrt(sP.sq)
> sP

+ [1] 340.8608

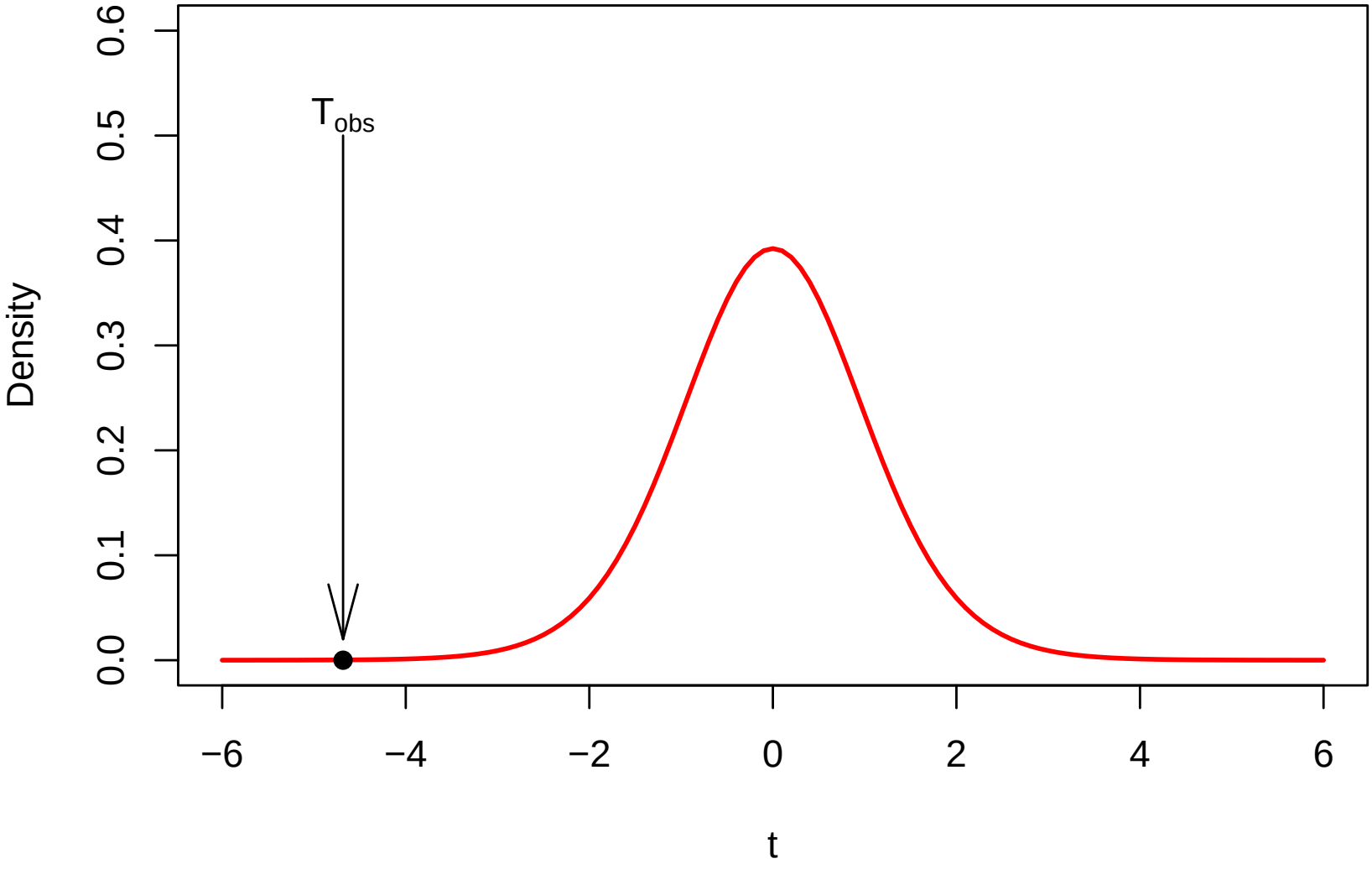
> Tobs<-(mean(xA)-mean(xB))/(sP*sqrt(1/nA+1/nB))
> Tobs

+ [1] -4.682925
```

Statistical testing



Statistical testing



Statistical testing

In this case, T_{obs} does not lie in a high probability region.

i.e. we are surprised to see T_{obs} so far away from zero.

The predicted behaviour of T assuming that H_0 is true

does not match the observed behaviour !

and so the assumption that H_0 is true must be incorrect. We

Reject H_0

in favour of H_a.

Statistical significance

How do we quantify the statistical significance ?

Two approaches:

1. Define the high probability region, and reject H_0 if T_{obs} does not lie in this region.

2. Compute the level of surprise at observing T_{obs} under the assumption that H_0 is true.

Statistical significance

For 1: Set significance level α, with $0 < \alpha < 1$, and find the central $1 - \alpha$ high probability region, between the two values $-C_R$ and C_R.

- α is the probability of a Type I Error, which occurs when we reject H_0, given that it is true

Statistical significance

- $\pm C_R$ are the critical values
  - here, C_R is the upper $\alpha/2$ point of the t_{15} distribution, denoted $t_{\alpha,15}$, where

$$P_0(T > t_{\alpha/2,15}) = \alpha/2.$$

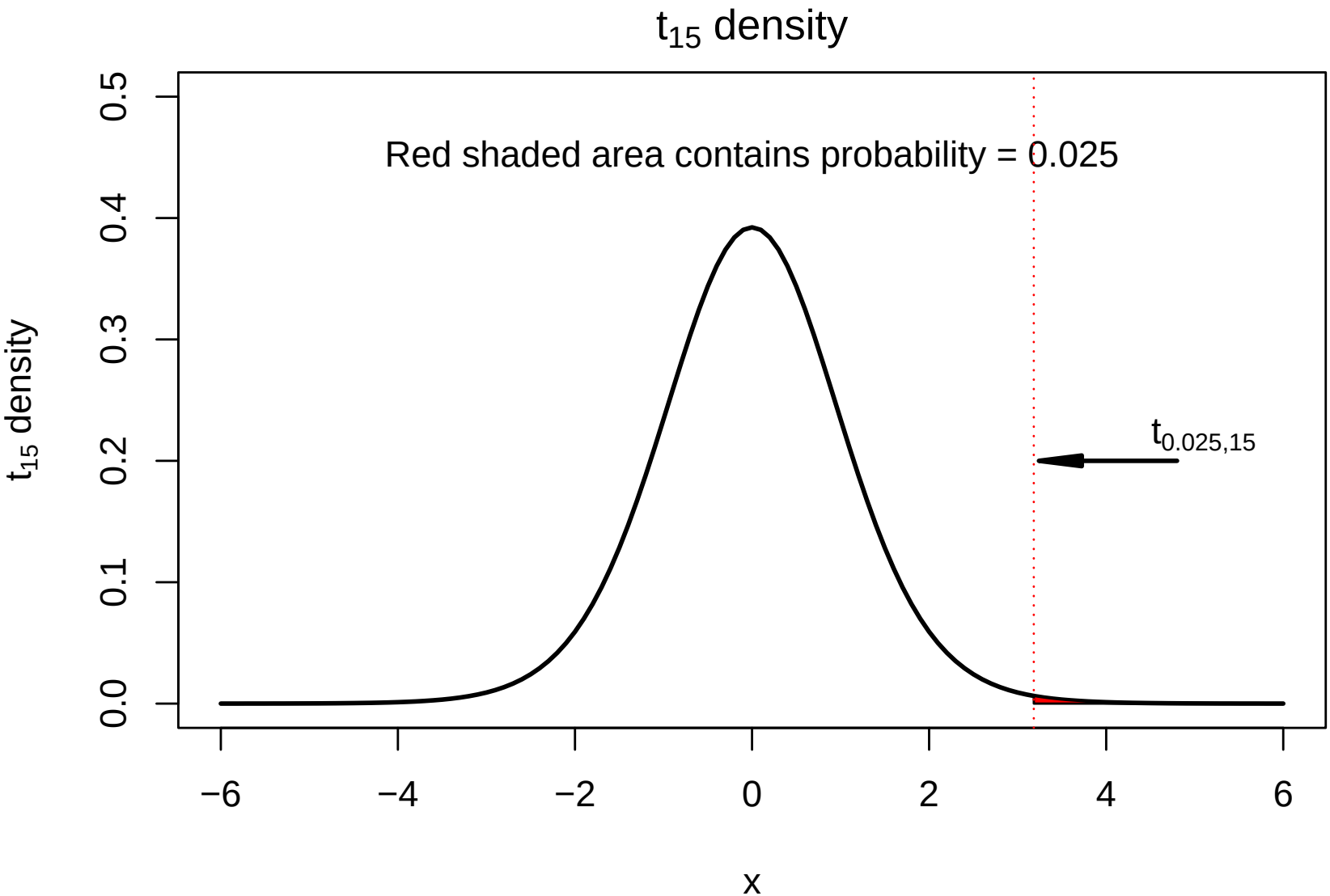
  - we use $\alpha/2$ as we are carrying out a two-tailed test
  - the total probability content of the high probability region is

$$1 - \alpha/2 - \alpha/2 = 1 - \alpha.$$

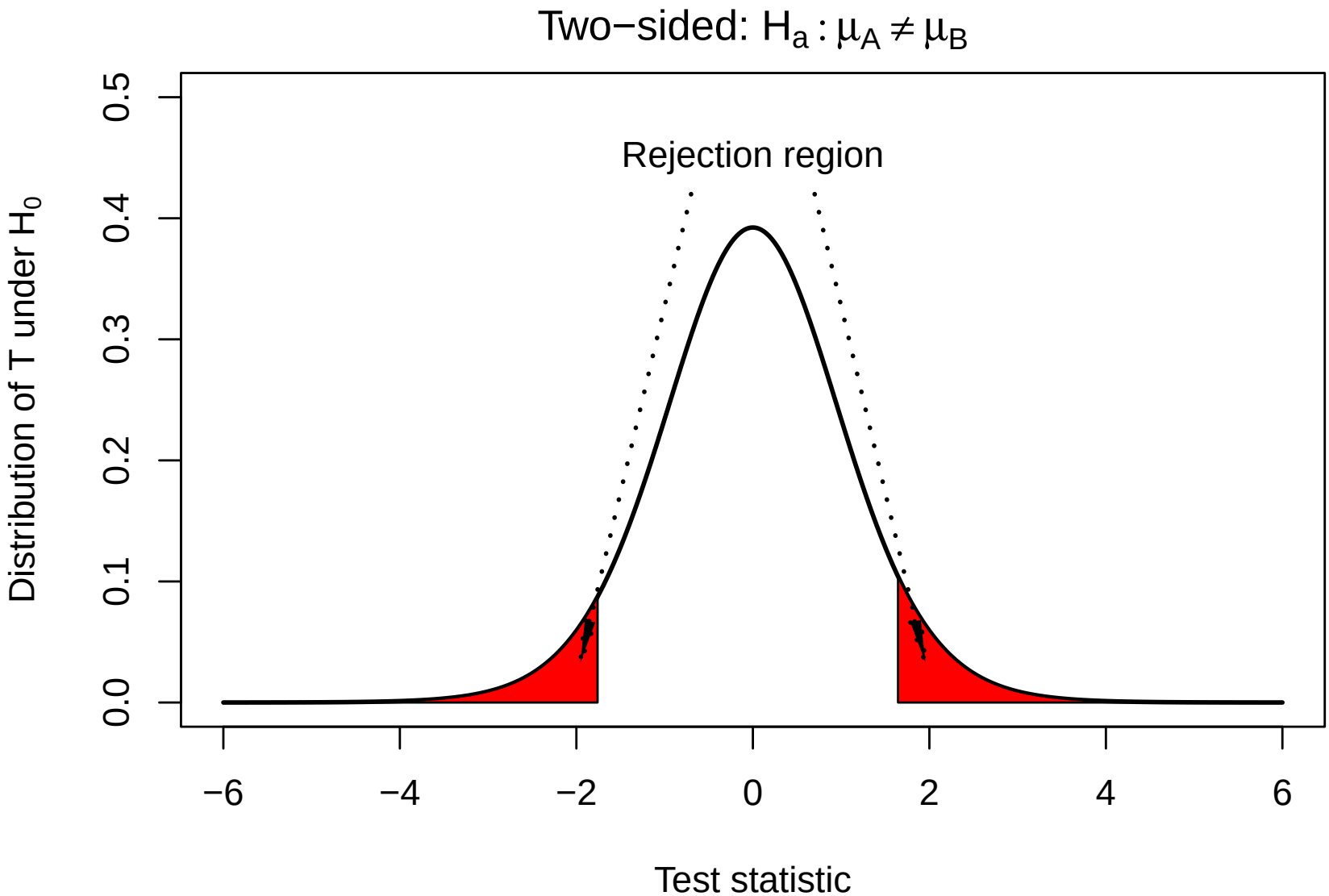
- the region outside the high probability region $(-C_R, C_R)$ is the rejection region $\mathcal{R}$.

If $T_{obs} \leqslant -C_R$ or $T_{obs} \geqslant C_R$, we reject H_0.

Statistical significance



Statistical significance



Statistical significance

Typically, $\alpha = 0.05$ (or 0.01), so for the t_{15} distribution

$$C_R = 2.131 \qquad (\text{or } 2.947)$$

As $T_{obs} = -4.683$ lies in the rejection region, we reject H_0.

Statistical significance

For 2: To compute the level of surprise, we evaluate the probability of observing a more extreme test statistic under the assumption that H_0 is true.

For the two-sided test, we report

$$2\Pr[T \geqslant |T_{obs}||H_0 \text{ is true}] = 2P_0(T \geqslant |T_{obs}|)$$

that is, in this case

$$P_0(T \leqslant -4.683) + P_0(T \geqslant 4.683)$$

which we can compute using tables or in software.

Statistical significance

Here, this probability is

$$p = 0.00029.$$

```
> Tobs

+ [1] -4.682925

> 2*(1-pt(abs(Tobs),15))    #P-value

+ [1] 0.0002944777
```

Statistical significance

This probability is very small, so we are very surprised by the observed result.

- p is termed the p-value.
- if $p \leqslant \alpha = 0.05$ (or 0.01), we reject H_0.

Confidence Intervals

Can also compute a 95 % confidence interval for $\mu_A - \mu_B$

$$(\overline{x}_A - \overline{x}_B) \pm t_{\alpha/2, n_A+n_B-2} s_P \sqrt{\frac{1}{n_A} + \frac{1}{n_B}}$$

where

$$t_{\alpha/2, n_A+n_B-2} \equiv t_{0.025,15} = 2.1314$$

is the upper 0.025 point of the t_{15} distribution.

Hence the 95 % confidence interval is

$$(-1144.59, -428.67)$$

which does not contain zero !

Further questions

Some general questions:

- How do we choose the test statistic ?
- How do we choose α ?
- Why is the distribution of T a t_{15} distribution ?
- How do we compute C_R and p ?

Equal Variances ?

Is the assumption of equal population variances

$$\sigma_A = \sigma_B$$

fair in this case ?

$$\begin{aligned} s_A^2 &= 61190.4 \\ s_B^2 &= 198679.6 \end{aligned}$$

so that

$$\frac{s_A^2}{s_B^2} = 0.3080.$$

Can we test $\sigma_A = \sigma_B$ formally ?

Equal Variances ?

Yes (see section 9.6 of textbook):

$$H_0: \quad \sigma_A = \sigma_B$$
$$H_a: \quad \sigma_A \neq \sigma_B$$

or equivalently

$$H_0: \quad \frac{\sigma_A^2}{\sigma_B^2} = 1$$

$$H_a: \quad \frac{\sigma_A^2}{\sigma_B^2} \neq 1$$

Equal Variances ?

Test statistic is

$$F = \frac{s_A^2}{s_B^2}$$

If H_0 is true, F should look like an observation from an

F-distribution

with

$$(n_A - 1, n_B - 1)$$

degrees of freedom.

Equal Variances ?

The F-distribution is just another probability distribution:

- the two degrees of freedom parameters determine the shape
- here, we have

$$n_A - 1 = 9 \qquad n_B - 1 = 6$$

so we need the $F_{9,6}$ distribution.

Equal Variances ?

Here the observed value of the test statistic is

$$F_{obs} = \frac{61190.400}{198679.619} = 0.308$$

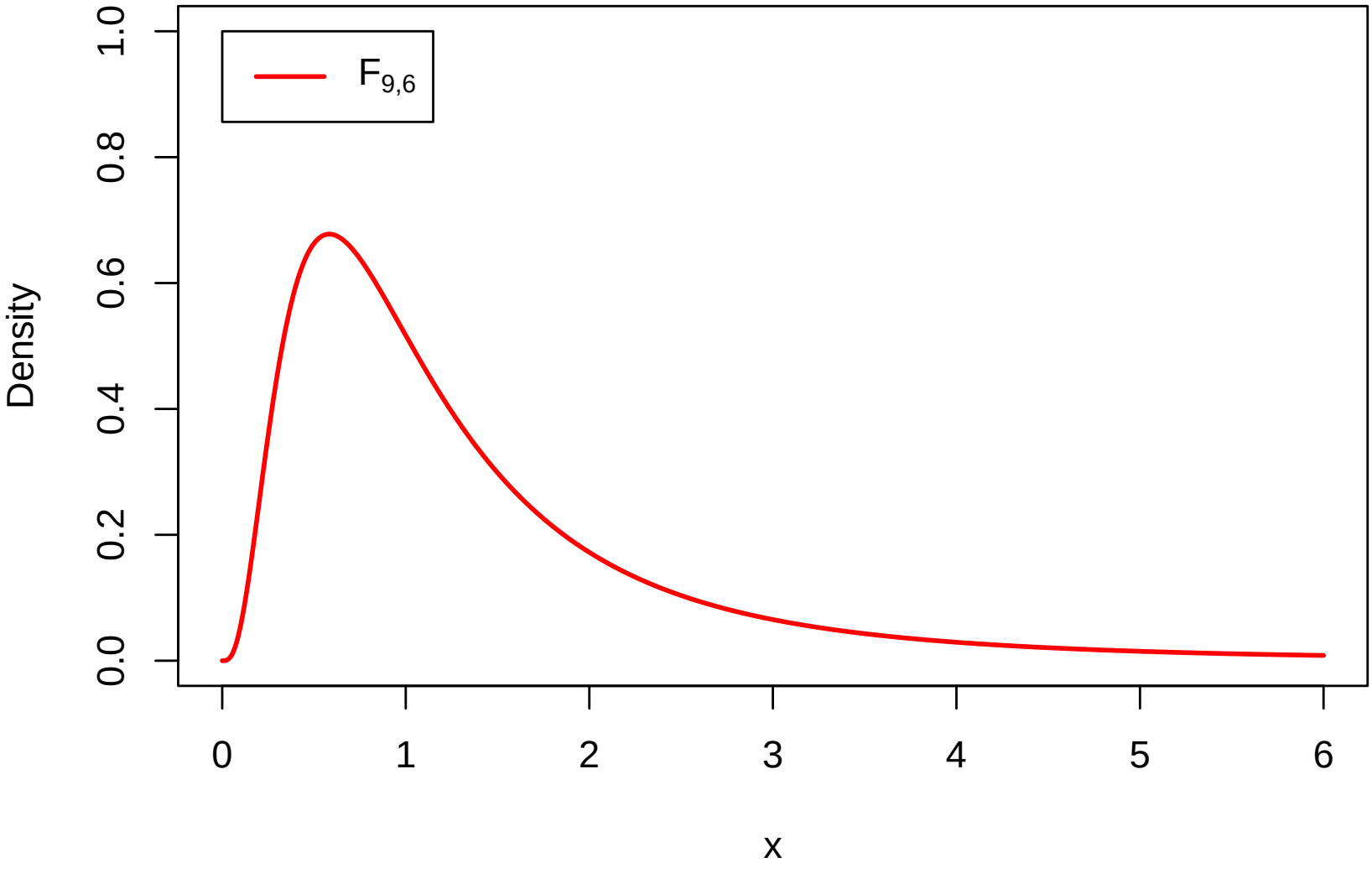
```
> Fval<-var(xA)/var(xB)
> Fval

+ [1] 0.3079853
```

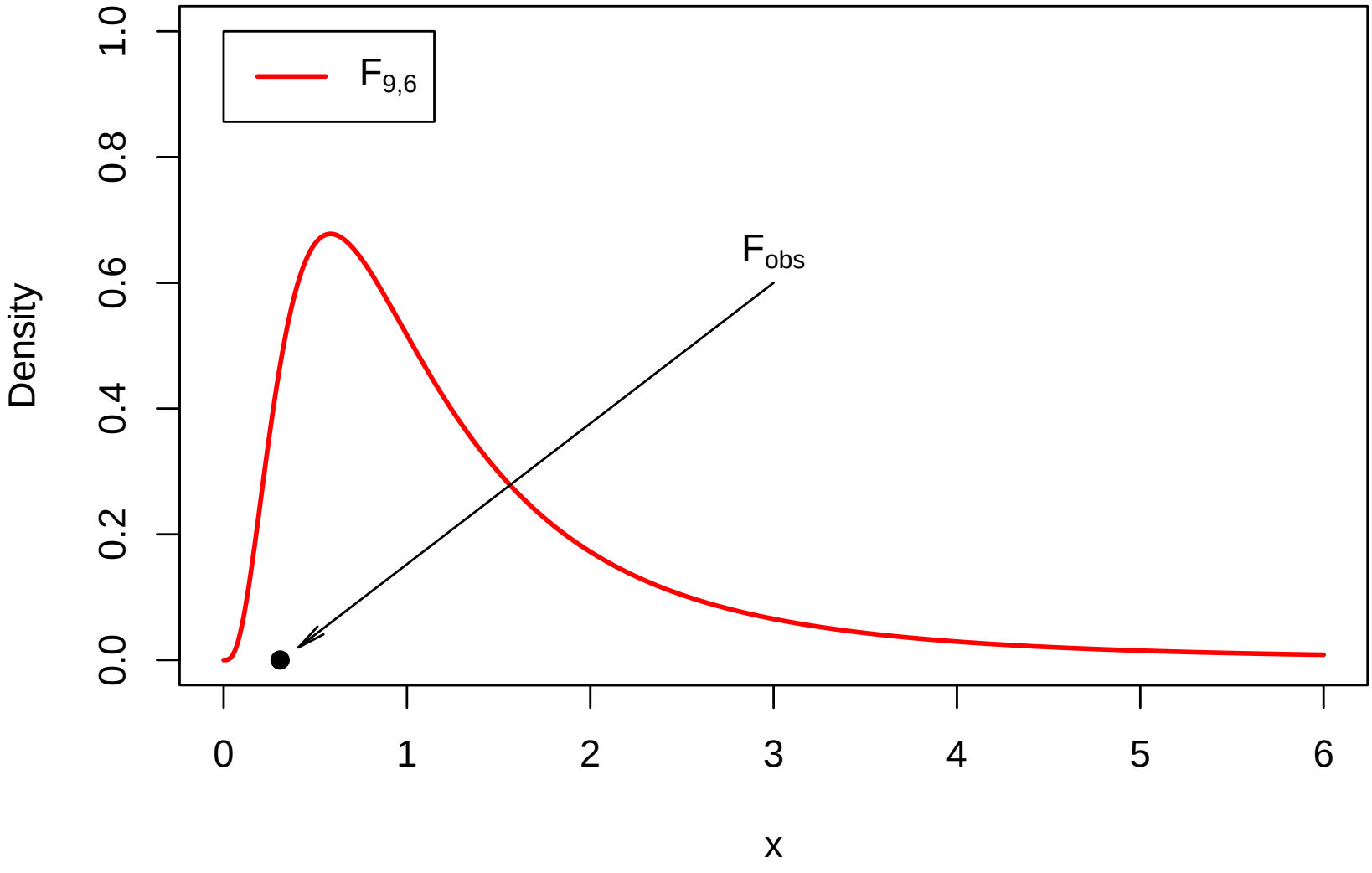
Values near 1 support H_0

- how far away can F_{obs} be before we start to prefer H_a ?
- we need to assess using the null distribution, here the $F_{9,6}$ distribution.

Equal Variances ?



Equal Variances ?



Equal Variances ?

Table: Upper $a = 0.05$ points of F_{ν_1,ν_2} distribution

	ν_1								
ν_2	1	2	3	4	5	6	7	8	9
1	161.45	199.50	215.71	224.58	230.16	233.99	236.77	238.88	240.54
2	18.51	19.00	19.16	19.25	19.30	19.33	19.35	19.37	19.38
3	10.13	9.55	9.28	9.12	9.01	8.94	8.89	8.85	8.81
4	7.71	6.94	6.59	6.39	6.26	6.16	6.09	6.04	6.00
5	6.61	5.79	5.41	5.19	5.05	4.95	4.88	4.82	4.77
6	5.99	5.14	4.76	4.53	4.39	4.28	4.21	4.15	4.10
7	5.59	4.74	4.35	4.12	3.97	3.87	3.79	3.73	3.68
8	5.32	4.46	4.07	3.84	3.69	3.58	3.50	3.44	3.39
9	5.12	4.26	3.86	3.63	3.48	3.37	3.29	3.23	3.18
10	4.96	4.10	3.71	3.48	3.33	3.22	3.14	3.07	3.02
11	4.84	3.98	3.59	3.36	3.20	3.09	3.01	2.95	2.90
12	4.75	3.89	3.49	3.26	3.11	3.00	2.91	2.85	2.80
13	4.67	3.81	3.41	3.18	3.03	2.92	2.83	2.77	2.71
14	4.60	3.74	3.34	3.11	2.96	2.85	2.76	2.70	2.65

From textbook, p 479. Reported to 2 dp.

Equal Variances ?

Note that this is the upper tail probability: that is,

$$P_0(F > F_{a,\nu_1,nu_2}) = a$$

and entries in the table are the numerical values denoted

$$F_{a,\nu_1,\nu_2}$$

for $a = 0.05$. For example,

$$F_{0.05,9,6} = 4.10$$

from the column headed 9, row headed 6.

Equal Variances ?

However, here we are doing a two-tailed test, so we need a rejection region that reflects this:

- if F_{obs} is very small or very large, this supports H_a !
- we need a lower tail region as well as an upper tail region
- we need to consider $a = \alpha/2 = 0.025$.

Equal Variances ?

Table: Upper $a = 0.025$ points of F_{ν_1,ν_2} distribution

	ν_1								
ν_2	1	2	3	4	5	6	7	8	9
1	647.79	799.50	864.16	899.58	921.85	937.11	948.22	956.66	963.28
2	38.51	39.00	39.17	39.25	39.30	39.33	39.36	39.37	39.39
3	17.44	16.04	15.44	15.10	14.88	14.73	14.62	14.54	14.47
4	12.22	10.65	9.98	9.60	9.36	9.20	9.07	8.98	8.90
5	10.01	8.43	7.76	7.39	7.15	6.98	6.85	6.76	6.68
6	8.81	7.26	6.60	6.23	5.99	5.82	5.70	5.60	5.52
7	8.07	6.54	5.89	5.52	5.29	5.12	4.99	4.90	4.82
8	7.57	6.06	5.42	5.05	4.82	4.65	4.53	4.43	4.36
9	7.21	5.71	5.08	4.72	4.48	4.32	4.20	4.10	4.03
10	6.94	5.46	4.83	4.47	4.24	4.07	3.95	3.85	3.78
11	6.72	5.26	4.63	4.28	4.04	3.88	3.76	3.66	3.59
12	6.55	5.10	4.47	4.12	3.89	3.73	3.61	3.51	3.44
13	6.41	4.97	4.35	4.00	3.77	3.60	3.48	3.39	3.31
14	6.30	4.86	4.24	3.89	3.66	3.50	3.38	3.29	3.21

Reported to 2 dp.

Equal Variances ?

We discover that,

$$F_{0.025,9,6} = 5.52$$

from the column headed 9, row headed 6. This handles the upper tail.

What about the lower tail ?

Equal Variances ?

It turns out that for the F-distribution we have the following result: if

$$X \sim F_{\nu_1,\nu_2}$$

and

$$Y = \frac{1}{X},$$

then

$$Y \sim F_{\nu_2,\nu_1}.$$

Equal Variances ?

This means that, for our test, for any x value

$$P_0(F \leqslant x) \equiv P_0(1/F \geqslant 1/x) \equiv P_0(1/F > 1/x)$$

where the symbol '$\equiv$' means 'is identical to'. In this expression, F is the variable representing the test statistic.

Equal Variances ?

But from the above result, we know that

$$\frac{1}{F} \sim F_{\nu_2,\nu_1}$$

which is enough information for us to compute the lower tail critical value.

Equal Variances ?

We

- Look up the $a = 0.025$ upper point of the $F_{6,9}$ distribution, and find out that

$$F_{0.025,6,9} = 4.32$$

  from the column headed 6, row headed 9.

- Take the reciprocal, to obtain that

$$F_{0.975,9,6} = \frac{1}{F_{0.025,6,9}} = \frac{1}{4.32} = 0.23$$

That is, the lower tail critical value is 0.23.

Equal Variances ?

We can compute more precisely using R:

```
> CR.1<-qf(0.025,9,6)
> CR.2<-qf(0.975,9,6)
> c(CR.1,CR.2)

+ [1] 0.2314964 5.5234066

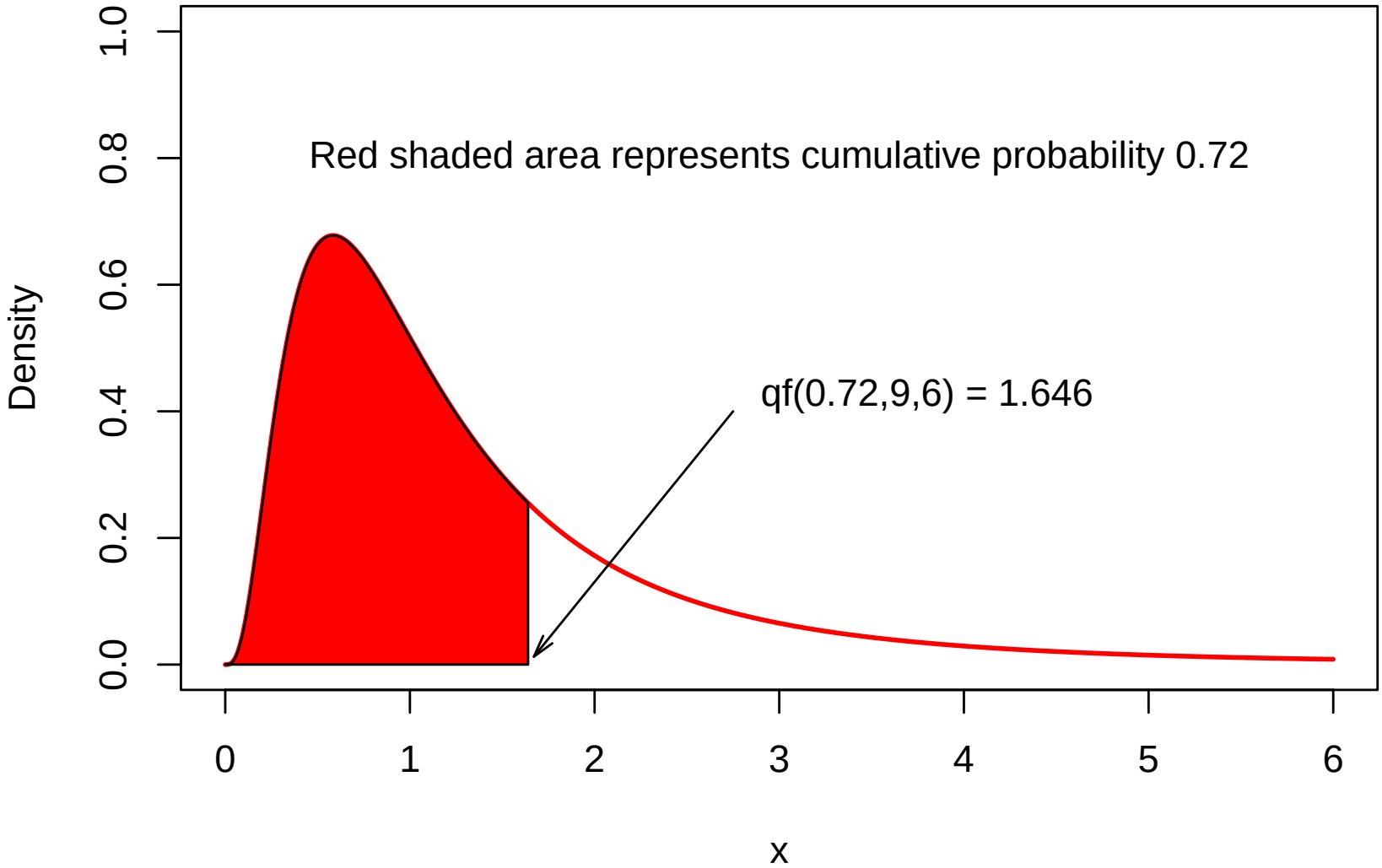
> CR.1<-1/qf(0.975,6,9)
> c(CR.1,CR.2)

+ [1] 0.2314964 5.5234066
```

The function `qf` takes three arguments

- the cumulative probability: for the upper a point, this is $1 - a$.
- the first degree of freedom, ν_1.
- the second degree of freedom, ν_2.

Equal Variances ?



Equal Variances ?

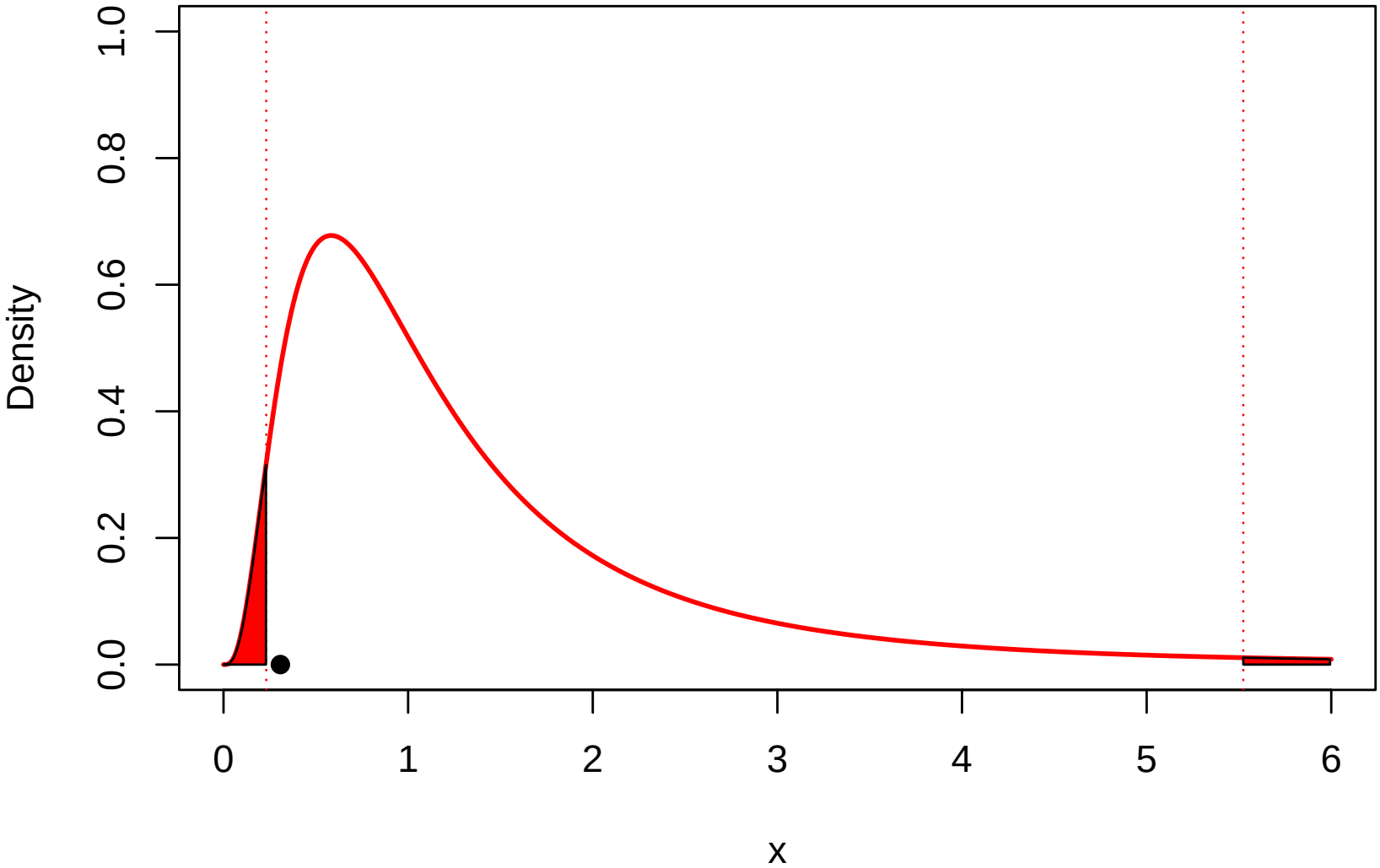
For $\alpha = 0.05$, the critical values in the test are

$$C_{R_1} = 0.231 \qquad C_{R_2} = 5.523$$

The observed value F_{obs} does lie in the high probability region, and there is no reason to reject H_0 at $\alpha = 0.05$.

For these data, there is no evidence that the population variances are different.

Equal Variances ?



Recap

This test of equal variances illustrates the general strategy for hypothesis testing:

- set up the hypotheses H_0 and H_a
- choose a suitable test statistic that allows assessment of the two hypotheses
- find the null distribution
- use the null distribution to define the rejection region
- compare the observed test statistic to the computed rejection region to complete the test

Course data sets

```
> e <- new.env(parent = emptyenv())
> load("204DataSets.rda", envir=e)
> ls(e)

+  [1] "AIDS"         "Batteries"    "Birthweight"  "Butterfat"    "Coleman"
+  [6] "Diabetes"     "DZT"          "Gravel"       "Haemo"        "HairEyes"
+ [11] "Hooker"       "Hypnosis"     "Lyrics"       "Memory"       "Milk"
+ [16] "MilkBac"      "Muco"         "PEFRAsthma"   "Potato"       "PTSD"
+ [21] "Pull"         "RBDExample"   "Sign1"        "Sign2"        "Soil"
+ [26] "TaskDistract" "Treadmill"    "Viscosity"
```

Birthweight data

```
> str(Birthweight)

+ 'data.frame': 17 obs. of  3 variables:
+  $ Group: Factor w/ 2 levels "A","B": 1 1 1 1 1 1 1 1 1 1 ...
+  $ gp   : int  1 1 1 1 1 1 1 1 1 1 ...
+  $ BW   : int  2164 2600 2184 2080 1820 2496 2184 2080 2184 2576 ...

> table(Birthweight$Group)

+
+  A  B
+ 10  7

> tapply(Birthweight$BW,Birthweight$Group,mean)

+        A        B
+ 2236.800 3023.429

> tapply(Birthweight$BW,Birthweight$Group,var)

+        A        B
+  61190.4 198679.6
```

Two-sample t-test

```
> t.test(BW~Group,data=Birthweight,var.equal=T)

+
+  Two Sample t-test
+
+ data:  BW by Group
+ t = -4.6829, df = 15, p-value = 0.0002945
+ alternative hypothesis: true difference in means between group A and group B is not equal to 0
+ 95 percent confidence interval:
+  -1144.6653  -428.5919
+ sample estimates:
+ mean in group A mean in group B
+        2236.800        3023.429
```

Test for equal variances

```
> var.test(BW~Group,data=Birthweight)

+
+  F test to compare two variances
+
+ data:  BW by Group
+ F = 0.30799, num df = 9, denom df = 6, p-value = 0.1106
+ alternative hypothesis: true ratio of variances is not equal to 1
+ 95 percent confidence interval:
+  0.05576002 1.33041078
+ sample estimates:
+ ratio of variances
+           0.3079853
```

Chapter 1
The Analysis of Variance

In this section

- Textbook Chapter 10
- introduction to the terminology of designed experiments
- extension of statistical testing theory to comparison of more than two population means
- The Analysis Of Variance (ANOVA) F-test

Designed Experiments

Studies typically fall into one of two categories:

(i) Observational studies: the experimenter has no control over the variables under study, and can only measure outcomes.

- The IQs of MAC and PC users
- The relationship between environmental exposure to toxins and health status.

i.e. The experimenter does not control the exposure to variables that may cause changes in the outcome of interest.

Designed Experiments

This type of study is common in medicine and epidemiology as it is relatively cheap to carry out.

Common type of observational study:

CASE-CONTROL STUDY

Designed Experiments

Example: Smoking and Lung Cancer

A study (Doll and Hill, 1950) investigated 649 lung cancer cases and 649 matched healthy controls, both drawn from a population of men in the UK. They found out what proportion in each group were smokers.

Neither health status nor smoking status were controlled by the experimenter, but were merely observed.

	Smokers	Non-smokers	Total
Lung cancer	647	2	649
Controls	622	27	649

Designed Experiments

This type of study can be unreliable, and cannot uncover all the relationships of interest.

A preferred approach involves the experimenter controlling the variables that cause variation in the other variables.

Note that this may not be ethical in a smoking/lung cancer study.

Designed Experiments

(ii) Designed experiments: the experimenter control the levels of variables that may affect the variable of interest.

For designed experiments, sometimes called experimental studies, the experimenter intervenes to assign the levels of influential variables such as

- dose levels of therapeutic treatments
- types of treatment

Designed Experiments

Example: Birthweight study

Recall

GROUP A : 5 or fewer visits

GROUP B : More than 5 visits.

If at the control of

(a) mothers, then this is an observational study

(b) doctors, this is a designed experiment

- for example, after each mother is recruited to take part in the study, they are randomly assigned to either GROUP A or GROUP B.

Designed Experiments

This is termed a

randomized experimental study

This type of study is preferable, but can be more difficult to implement.

Terminology

- Response variable (dependent variable):
  - the variable of interest in the study
- Factors : the variables that may have an effect on the response variable
  - quantitative if measured on a numerical scale
  - qualitative otherwise

Terminology

- Factor Levels: the values of the factors utilized in the experiment
- Treatments: the factor-level combinations utilized.
- Experimental Units (subjects):
  - the objects on which the factors are measured or observed.

Terminology

Therefore:

- A designed experiment is one for which the analyst or experimenter controls the specification of treatments and the method of assigning units to treatments.

- An observational experiment or study is one for which the analyst simply observes the treatments and response on a sample of experimental units.

Terminology

Example: Birthweight study

- Response: Birthweight (g)
- Factor: Pre-natal treatment group
- Factor levels: GROUP A or GROUP B

that is, we have a single factor with two factor levels.

Terminology

Example: Test scores

Test scores of students in four schools for two types of test.

- Response: Test score
- Factors: TEST and SCHOOL (both qualitative)
- Factor levels:
  - TEST: 1, 2
  - SCHOOL: A, B, C, D

that is, we have a two factors, TEST with two factor levels and SCHOOL with four factor levels.

There are 8 possible treatments:

$$(1, A), (1, B), (1, C), (1, D), (2, A), (2, B), (2, C), (2, D)$$

Terminology



https://cihr-irsc.gc.ca/e/50833.html

Terminology

Example: Pain Relief

Different pain relief remedies are to be compared.

- REMEDY (quantitative/qualitative, 3 levels)
  - Dose levels 0, 1, 2
- AGE GROUP (quantitative/qualitative, 4 levels)
  - 0-16 years
  - 17-40 years
  - 41-65 years
  - 66 years and over
- SEX (qualitative, 2 levels)
  - Female
  - Male

Terminology

In this example, there are a total of $3 \times 4 \times 2 = 24$ possible treatments;

- REMEDY is the only factor that can be assigned by the analyst
- We might consider stratification into the 4×2 strata defined by the other two factors.

The Completely Randomized Design

A completely randomized design (CRD) is a design for which treatments are randomly assigned to experimental units, or in which random samples of experimental units are selected for each treatment.

- if the 'treatments' are DOSE 0/DOSE 1, a CRD assigns experimental units independently to the two treatment groups at random.

Statistical Objectives

The experimental units assigned to different treatments (that is, factor-level combinations) form

independent samples

from

different populations

in a CRD.

We wish to compare treatments: specifically, we wish to compare the treatment-specific population means.

A Multiple Group Comparison of Means !

Comparing k Treatments

Suppose that there are k treatments:

TMT 1	Mean μ_1
TMT 2	Mean μ_2
$\vdots$	$\vdots$
TMT k	Mean μ_k

We wish to test the hypotheses

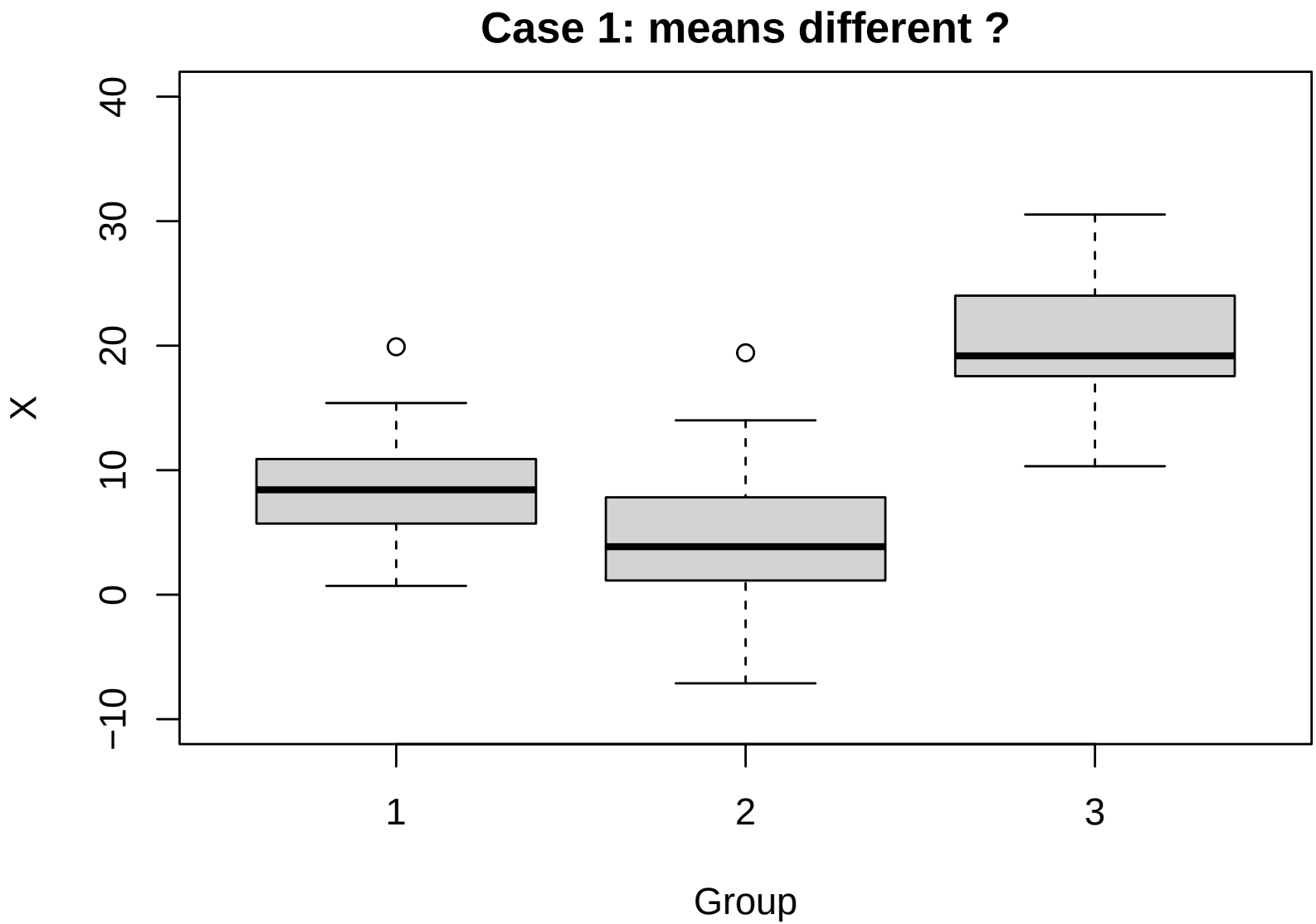
H_0 : $\mu_1 = \mu_2 = \cdots = \mu_k$

H_a : At least two of the k treatment means are different

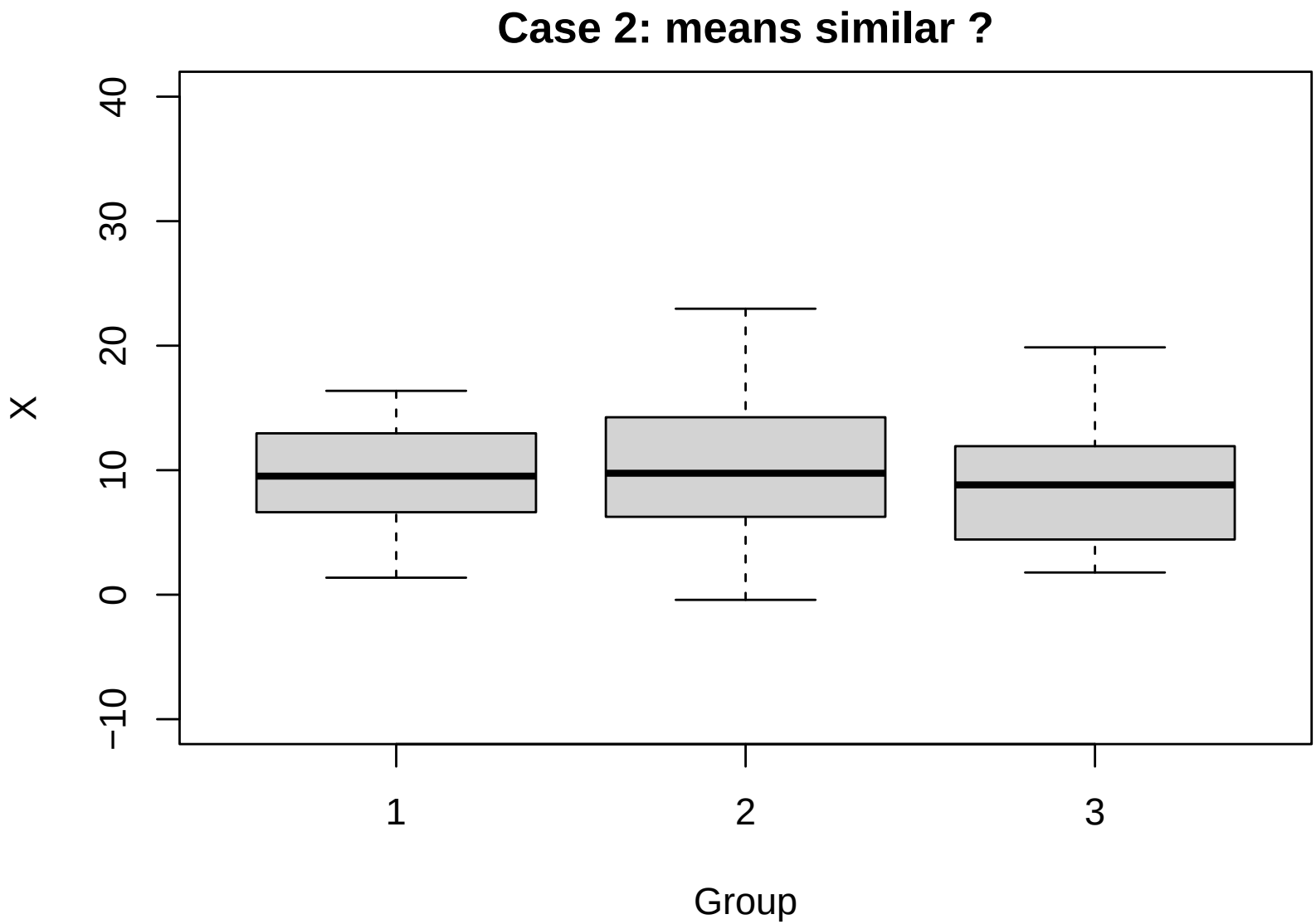
How do we do this ?

What is the relevant test statistic ?

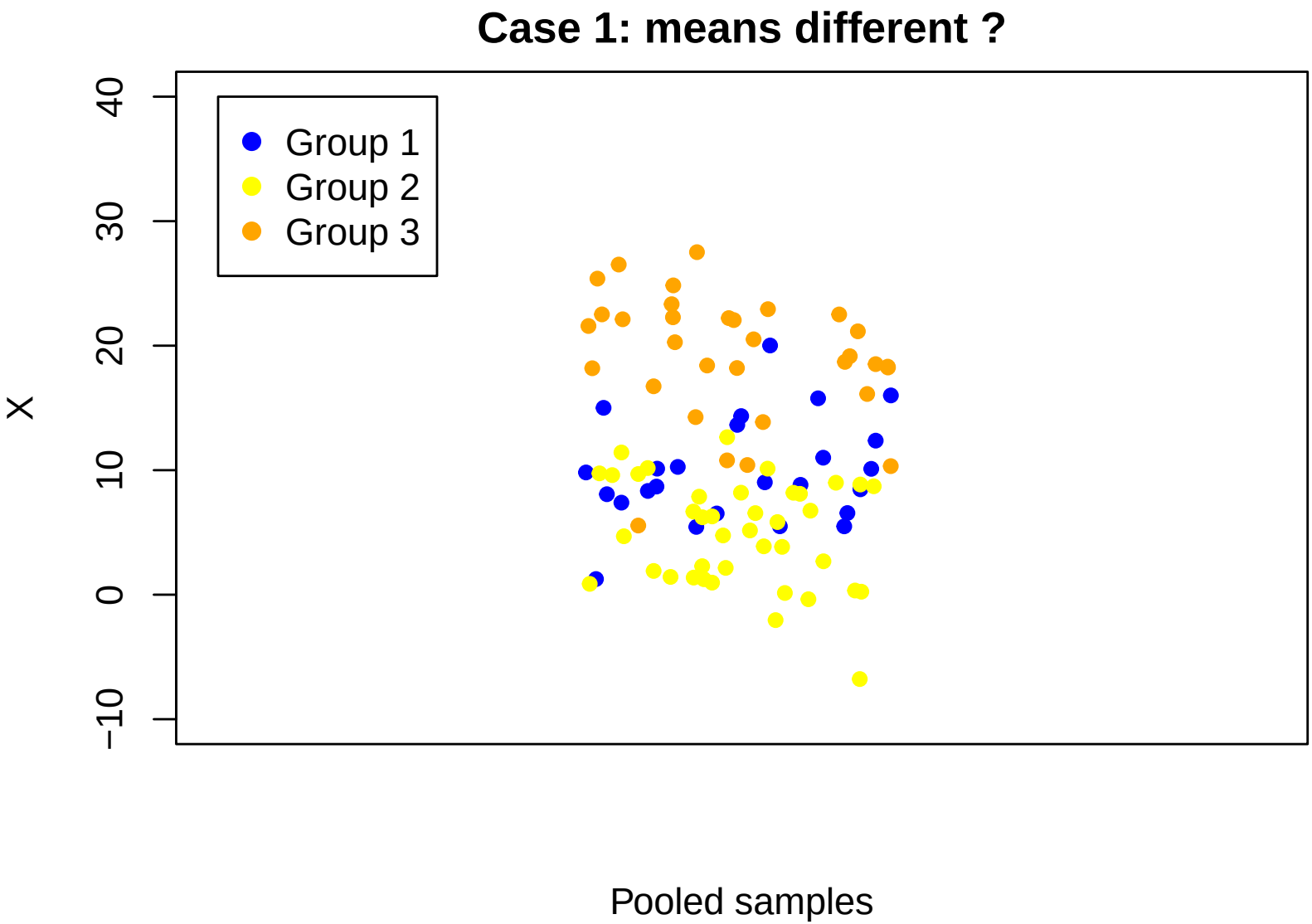
Comparing k Treatments



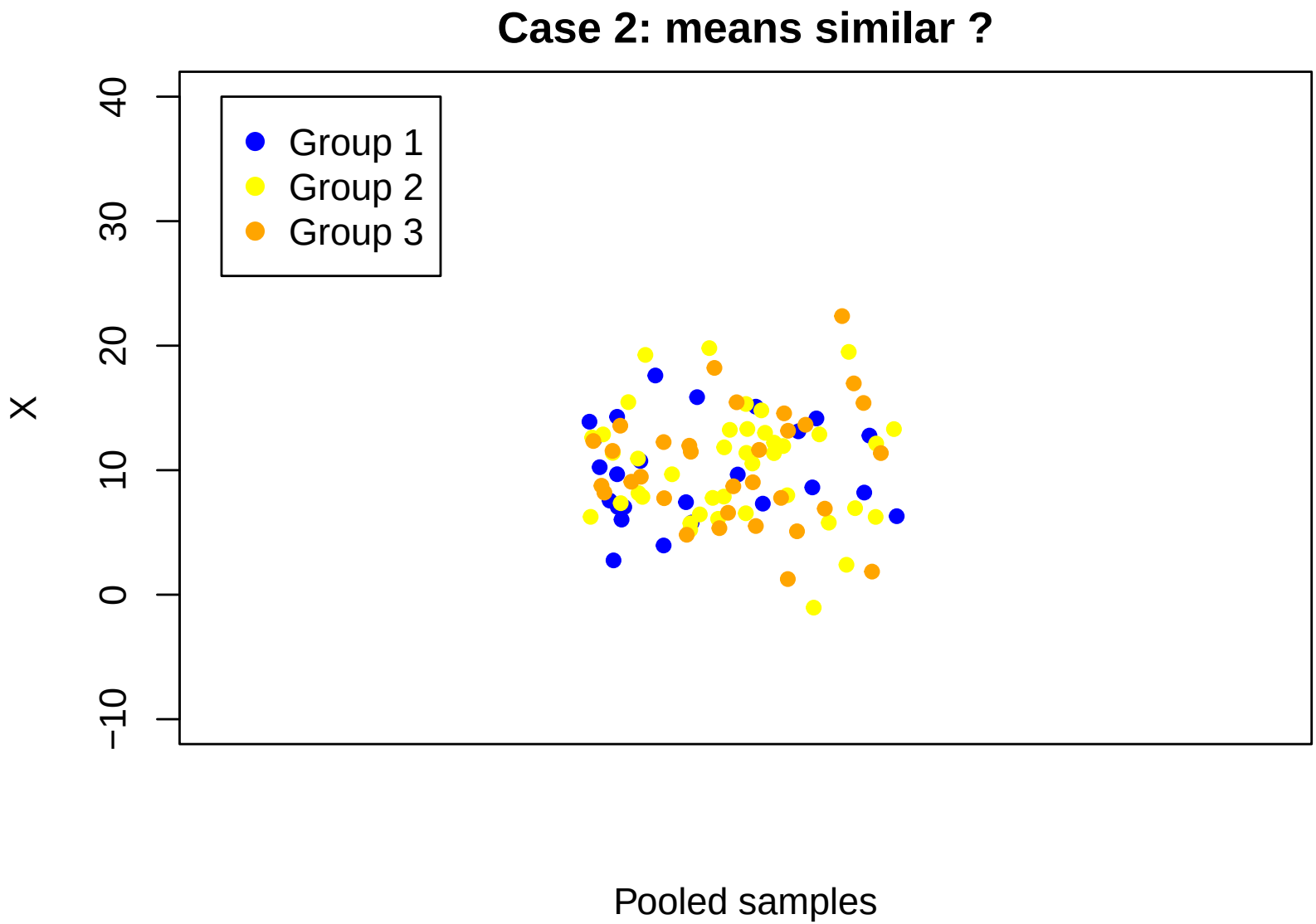
Comparing k Treatments



Comparing k Treatments



Comparing k Treatments



Comparing k Treatments

We want to find a way to compare

the between group variability

with the

the overall variability

we observe in the data.

How much of the overall variability is due to the fact that the means are different ?

The ANOVA F-test for CRDs

We attempt to generalize the earlier two sample comparisons:

- find a suitable test statistic
- predict how it should behave if H_0 is true
- compute the observed value of the test statistic
- compare the observed value to the rejection region.

The ANOVA F-test for CRDs

Suppose

$$\begin{array}{ll} \text{TMT } 1 & \text{has } n_1 \text{ experimental units} \\ \text{TMT } 2 & \text{has } n_2 \text{ experimental units} \\ \vdots & \vdots \\ \text{TMT } k & \text{has } n_k \text{ experimental units} \end{array}$$

Denote by x_{ij} the response for unit j in treatment group i, for $j = 1, \ldots, n_i$ and $i = 1, \ldots, k$.

The ANOVA F-test for CRDs

Our model for how the data are generated is

$$X_{ij} = \mu_i + \epsilon_{ij}$$

where ϵ_{ij} are independent residual errors.

We typically assume that $\epsilon_{ij} \sim Normal(0, \sigma_i^2)$, so that

$$X_{ij} \sim Normal(\mu_i, \sigma_i^2).$$

The ANOVA F-test for CRDs

Let

$$\overline{x}_i = \frac{1}{n_i} \sum_{j=1}^{n_i} x_{ij}$$

denote the sample mean for treatment i, and

$$s_i^2 = \frac{1}{n_i - 1} \sum_{j=1}^{n_i} (x_{ij} - \overline{x}_i)^2$$

denote the sample variance for treatment i.

Now we consider pooling, that is, combining all units into a single group.

The ANOVA F-test for CRDs

Define

- the total sample size

$$n = n_1 + \cdots + n_k = \sum_{i=1}^{k} n_i$$

- the overall sample mean

$$\overline{x} = \frac{1}{n} \sum_{i=1}^{k} \sum_{j=1}^{n_i} x_{ij}$$

- the overall sample variance

$$s^2 = \frac{1}{n-1} \sum_{i=1}^{k} \sum_{j=1}^{n_i} \left(x_{ij} - \overline{x}\right)^2$$

The ANOVA F-test for CRDs

Finally, consider the pooled sample variance

$$s_P^2 = \frac{1}{n-k} \sum_{i=1}^{k} (n_i - 1) s_i^2$$

This is the extension of the pooled estimate of the population variance in a two-sample t-test.

Using these quantities, we can derive a test statistic for multiple group comparison.

The ANOVA F-test for CRDs

We wish to compare how much variation is due to the

A difference between treatments

and how much is due to

B random variation within treatments

The ANOVA F-test for CRDs

We measure $\boxed{\text{A}}$ using the statistic

$$SST = \sum_{i=1}^{k} n_i (\overline{x}_i - \overline{x})^2$$

SST - Sum of Squares for Treatments

The ANOVA F-test for CRDs

We measure B using the statistic

$$SSE = \sum_{i=1}^{k} \sum_{j=1}^{n_i} (x_{ij} - \overline{x}_i)^2 = (n-k)s_P^2$$

SSE - Sum of Squares for Error

The ANOVA F-test for CRDs

NOTE: This measure of random or error variability implicitly assumes that the variability within the treatment groups is the same for each group. That is, population variances

$$\sigma_1^2, \ldots, \sigma_k^2$$

are equal, so that

$$X_{ij} \sim Normal(\mu_i, \sigma^2)$$

for all i and j.

In practice this assumption must be checked.

The ANOVA F-test for CRDs

We define the test statistic using the average levels of variability:

- MST - Mean Square for Treatments

$$MST = \frac{SST}{k-1} = \frac{1}{k-1}\sum_{i=1}^{k} n_i\left(\overline{x}_i - \overline{x}\right)^2$$

- MSE - Mean Square for Error

$$MSE = \frac{SSE}{n-k} = \frac{1}{n-k}\sum_{i=1}^{k}\sum_{j=1}^{n_i}\left(x_{ij} - \overline{x}_i\right)^2 = s_P^2$$

The ANOVA F-test for CRDs

Then the test statistic is

$$F = \frac{MST}{MSE} = \frac{\text{Average Variation due to Treatments}}{\text{Average Variation due to Errors}}$$

F large $\Longrightarrow$ Treatments Different !
F small $\Longrightarrow$ Treatments Similar !

The behaviour of F is given by the following Theorem.

The ANOVA F-test for CRDs

Theorem: ANOVA F-test to compare k treatments in a Completely Randomized Design

To test the hypothesis of equal treatment means,

$$H_0 \quad : \quad \mu_1 = \mu_2 = \cdots = \mu_k$$

$$H_a \quad : \quad \text{At least two of the } k \text{ treatment means are different}$$

the test statistic is

$$F = \frac{MST}{MSE}$$

The ANOVA F-test for CRDs

Theorem: ANOVA F-test to compare k treatments in a Completely Randomized Design

If H_0 is true, then

$$F \sim F_{k-1,n-k}$$

and the rejection region for a test at significance level α is the region to the right of the upper α point of this F distribution.

The ANOVA F-test for CRDs

Note: The ANOVA F-test is a one-tailed test – we reject H_0 if

$$F_{obs} \geqslant F_{\alpha,k-1,n-k}$$

- large values of F_{obs} support H_a
- small values of F_{obs} support H_0.

The ANOVA F-test for CRDs

Note: If

$$SS = \sum_{i=1}^{k} \sum_{j=1}^{n_i} (x_{ij} - \overline{x})^2$$

is the overall or total sum of squares, then we have that

$$SS = SST + SSE$$

so we can decompose the overall variation (SS) into the variation due to treatments (SST) and the variation due to the errors (SSE).

Assumptions behind the ANOVA F-test

There are three assumptions:

1. The samples are randomly selected in an independent manner from the k treatment populations.

2. All k treatment groups have distributions that are Normal (at least approximately).

3. The k treatment group variances are equal:

$$\sigma_1^2 = \sigma_2^2 = \cdots = \sigma_k^2.$$

Example

Example: Milk Quality Data

The impact on milk protein level of three different diets is being studied.

Data: Measurements of protein levels for $n = 1337$ samples.

- Response: Milk Protein Level (%)
- Factor: DIET
- Factor levels: $k = 3$
  - 1: Barley
  - 2: Barley + Lupins
  - 3: Lupins

Example

Data summary:

	TMT 1	TMT 2	TMT 3
n_i	425	459	453
$\overline{x}_i$	3.532	3.430	2.312
s_i^2	0.102	0.091	0.114

$$\begin{aligned} SST &= 10.606 \\ SSE &= 136.432 \\ SS &= 147.038 \end{aligned}$$

$$\begin{aligned} k-1 &= 2 \\ n-k &= 1334 \end{aligned}$$

Example

Therefore

$$MST = \frac{SST}{k-1} = \frac{10.606}{2} = 5.303$$

$$MSE = \frac{SSE}{n-k} = \frac{136.432}{1334} = 0.102$$

and

$$F_{obs} = \frac{MST}{MSE} = 51.851$$

Example

If H_0 is true, that is,

$$\mu_1 = \mu_2 = \mu_3$$

then F should look like an observation from a

$$F_{k-1,n-k}$$

distribution. Here we are dealing with the

$$F_{2,1334}$$

distribution.

Example

From R, we discover that if $\alpha = 0.05$, then the relevant upper point is

$$F_{0.05,2,1334} = 3.002$$

and thus we

$$\text{Reject } H_0$$

and conclude that there is a significant impact on milk protein level due to diet.

Example

Note: Table VI in the textbook (p 822) only gives

$$\begin{aligned} F_{0.05,2,120} &= 3.07 \\ F_{0.05,2,\infty} &= 3.00 \end{aligned}$$

so we cannot look up $F_{0.05,2,1334}$ exactly. However, we do know that

$$3.00 < F_{0.05,2,1334} < 3.07$$

and here the test statistic is $F_{obs} = 51.851$.

Are the assumptions met ?

1. Independent samples ? Not possible to tell with current information. In fact, data comprise repeated measurements on 79 cows - potentially not independent, as observations on the same cow are likely to be more similar.

2. Normal Distributions ? Visual inspection of boxplots indicates that this may be valid.

3. Equal variances ? : group sample variances are

$$s_1^2 = 0.102 \qquad s_2^2 = 0.091 \qquad s_3^2 = 0.114$$

   so assumption appears to be valid.

Milk data: Analysis in R

```
> head(Milk)

+   protein Time Cow   Diet
+ 1    3.63    1 B01 barley
+ 2    3.57    2 B01 barley
+ 3    3.47    3 B01 barley
+ 4    3.65    4 B01 barley
+ 5    3.89    5 B01 barley
+ 6    3.73    6 B01 barley
```

Milk data: Analysis in R

Summary statistics by treatment group:

```
> tapply(Milk$protein,Milk$Diet,length)

+        barley barley+lupins        lupins
+           425           459           453

> tapply(Milk$protein,Milk$Diet,mean)

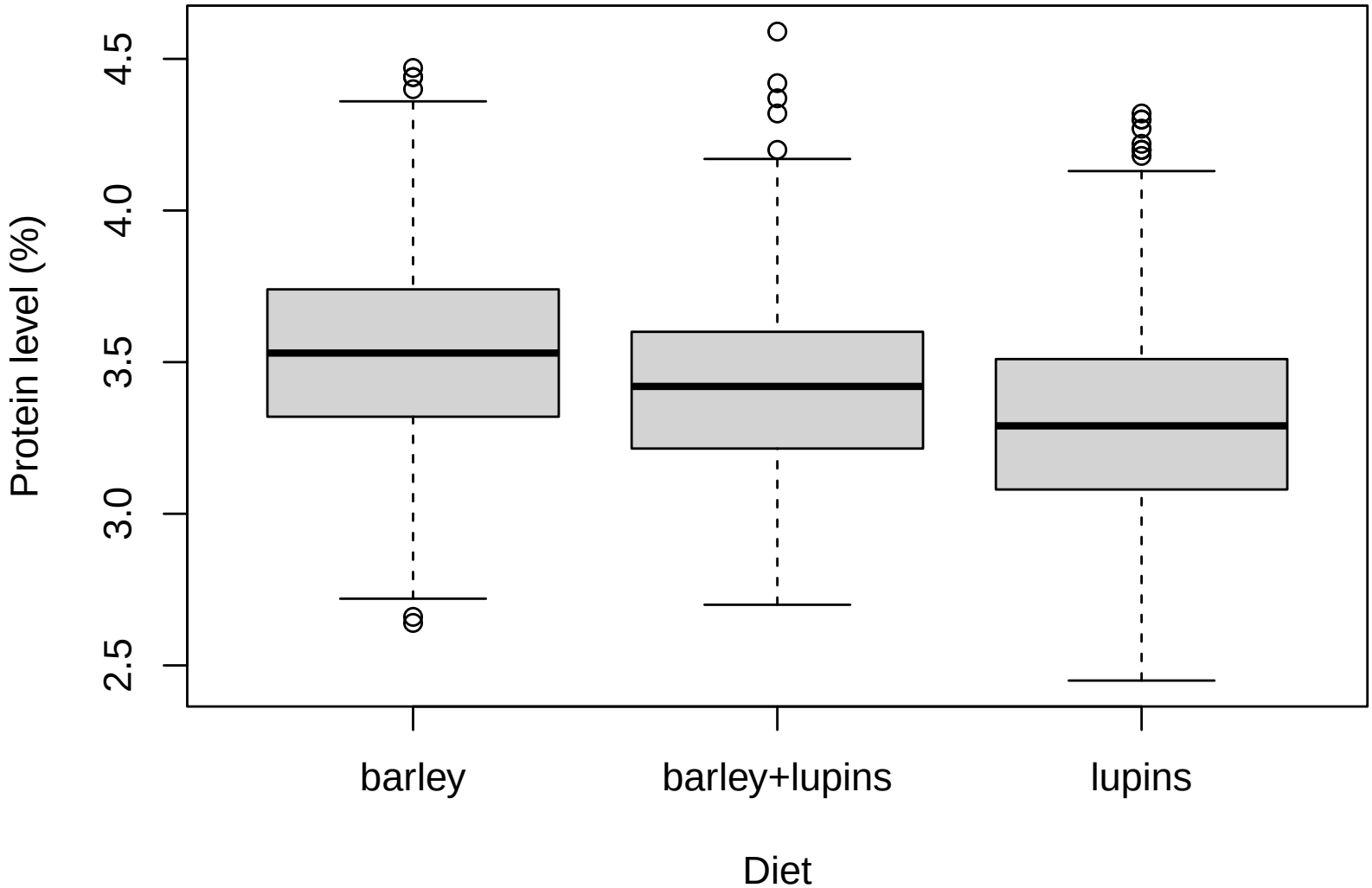
+        barley barley+lupins        lupins
+      3.531929      3.429695      3.312384

> tapply(Milk$protein,Milk$Diet,var)

+        barley barley+lupins        lupins
+    0.10189721    0.09141126    0.11363103

> #Boxplot
> par(mar=c(4,4,2,2))
> boxplot(protein~Diet,data=Milk, ylab='Protein level (%)')
```

Milk data: Analysis in R



Milk data: Analysis in R

Compute the total sum of squares (SS):

```
> xbar<-mean(Milk$protein)
>
> SS<-sum((Milk$protein-xbar)^2)
> SS

+ [1] 147.0379
```

Milk data: Analysis in R

Compute the sum of squares for treatments (SST):

```
> nvec<-tapply(Milk$protein,Milk$Diet,length)
> mvec<-tapply(Milk$protein,Milk$Diet,mean)
> SST<-sum(nvec*(mvec-xbar)^2)
> SST

+ [1] 10.6059
```

Milk data: Analysis in R

Compute the sum of squares for error (SSE):

```
> svec<-tapply(Milk$protein,Milk$Diet,var)
> SSE<-sum((nvec-1)*svec)
> SSE

+ [1] 136.432
```

Note that

$$SS = SST + SSE$$

that is

$$147.0379023 = 10.605902 + 136.4320003$$

Milk data: Analysis in R

Using aov and anova functions

```
> summary(aov(protein~1,data=Milk))

+               Df Sum Sq Mean Sq F value Pr(>F)
+ Residuals   1336    147  0.1101

> anova(lm(protein~1,data=Milk))

+ Analysis of Variance Table
+
+ Response: protein
+             Df Sum Sq Mean Sq F value Pr(>F)
+ Residuals 1336 147.04 0.11006
```

Milk data: Analysis in R

The result of the F-test is also computed

```
> summary(aov(protein~Diet,data=Milk))

+               Df Sum Sq Mean Sq F value Pr(>F)
+ Diet           2  10.61   5.303   51.85 <2e-16 ***
+ Residuals   1334 136.43   0.102
+ ---
+ Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

Milk data: Analysis in R

Look up the critical value using the `qf` function, and compute the p-value using the `pf` function.

```
> qf(0.95,2,1334)

+ [1] 3.00247

> 1-pf(51.85,2,1334)

+ [1] 0
```

Note that the `pf` function computes the cumulative probability (that is, in the left-tail of the distribution); for the p-value, we need the right-tail probability.

Milk data: Analysis in R

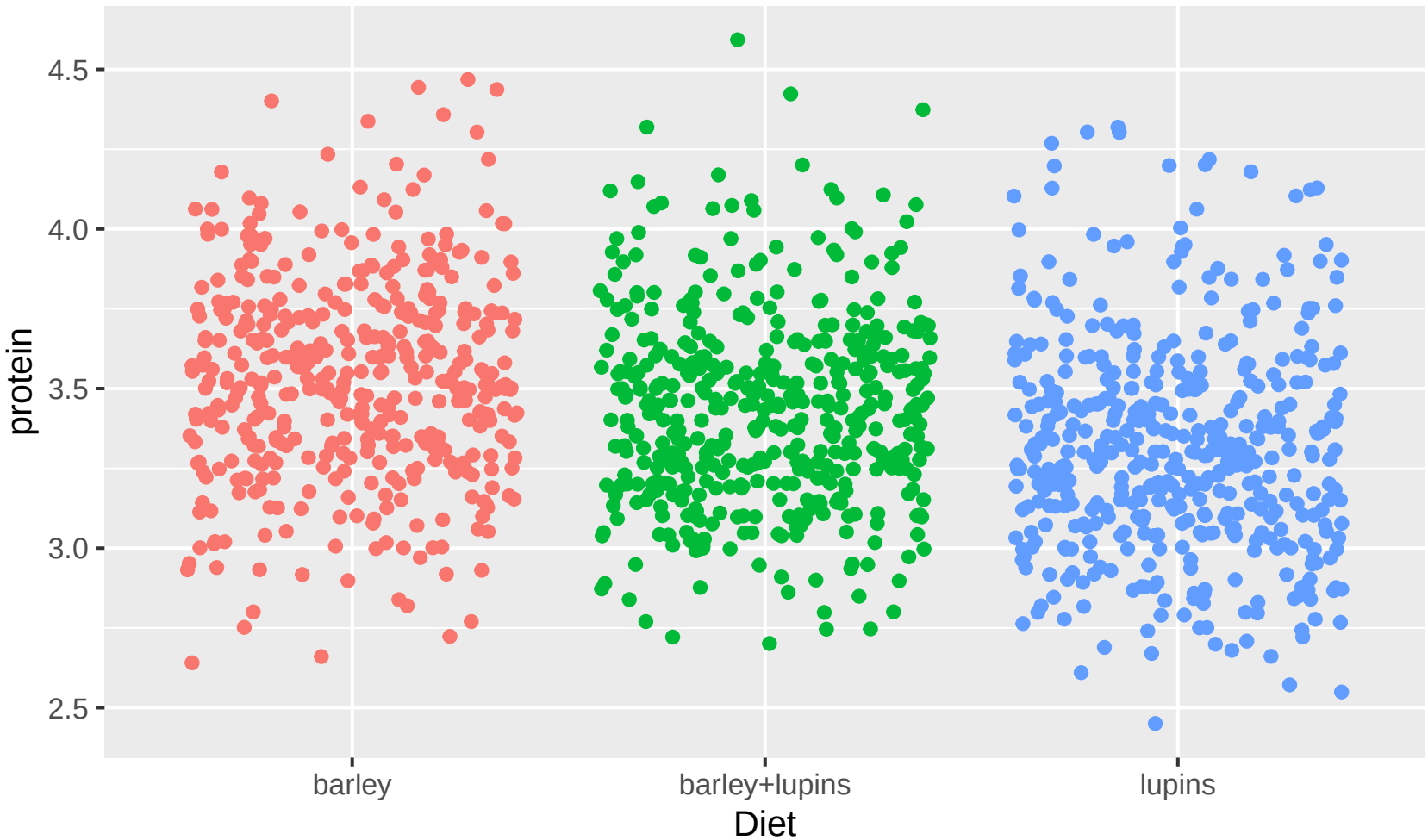
```
> library(supernova)
> supernova(lm(protein~Diet,data=Milk))

+  Analysis of Variance Table (Type III SS)
+  Model: protein ~ Diet
+
+                             SS   df    MS      F   PRE     p
+  ----- --------------- | ------- ---- ----- ------ ----- -----
+  Model (error reduced) |  10.606    2 5.303 51.851 .0721 .0000
+  Error (from model)    | 136.432 1334 0.102
+  ----- --------------- | ------- ---- ----- ------ ----- -----
+  Total (empty model)   | 147.038 1336 0.110
```

Milk data: Analysis in R

```
> library(ggplot2)
> ggplot(Milk) +
+   aes(x = Diet, y = protein, color = Diet) +
+   geom_jitter() +
+   theme(legend.position = "none")
```

Milk data: Analysis in R



Example: Anxiety Response

Example: Anxiety Response Treatment

In a study of Alzheimer's disease and care of its sufferers, a medication designed to improve anxiety relief has been developed.

In a lab experiment, $n = 20$ rats were assigned to one of four ($k = 4$) treatment groups corresponding to dose-level of the medication.

A measure of response to a "flee stimulus" was recorded.

Example: Anxiety Response

Example: Anxiety Response Treatment

- Response: Pull response to stimulus (units of force)
- Factor: DOSE-LEVEL
- Factor levels: $k = 4$
  - Dose 0 (zero units)
  - Dose 1 (one unit)
  - Dose 2 (two units)
  - Dose 3 (three units)

Example: Anxiety Response

Data:

0	1	2	3
27.0	22.8	21.9	23.5
26.2	23.1	23.4	19.6
28.8	27.7	20.1	23.7
33.5	27.6	27.8	20.8
28.8	24.0	19.3	23.9

We have $k = 4$ and $n = 20$, so

$$k - 1 = 3 \qquad n - k = 16.$$

Example: Anxiety Response

We find that

$$SST = 140.094 \qquad SSE = 116.324 \qquad SS = 256.418$$

$$MST = 46.698 \qquad MSE = 7.270$$

and

$$F_{obs} = 6.423$$

which we need to compare with the $F_{3,16}$ distribution.

Example: Anxiety Response

For $\alpha = 0.05$, from textbook tables

$$F_{0.05,3,16} = 3.24$$

and so we

$$\text{Reject } H_0$$

at $\alpha = 0.05$ and conclude that there is a significant difference between treatment groups.

p-value is 0.0046.

Example: Anxiety Response

Summaries:

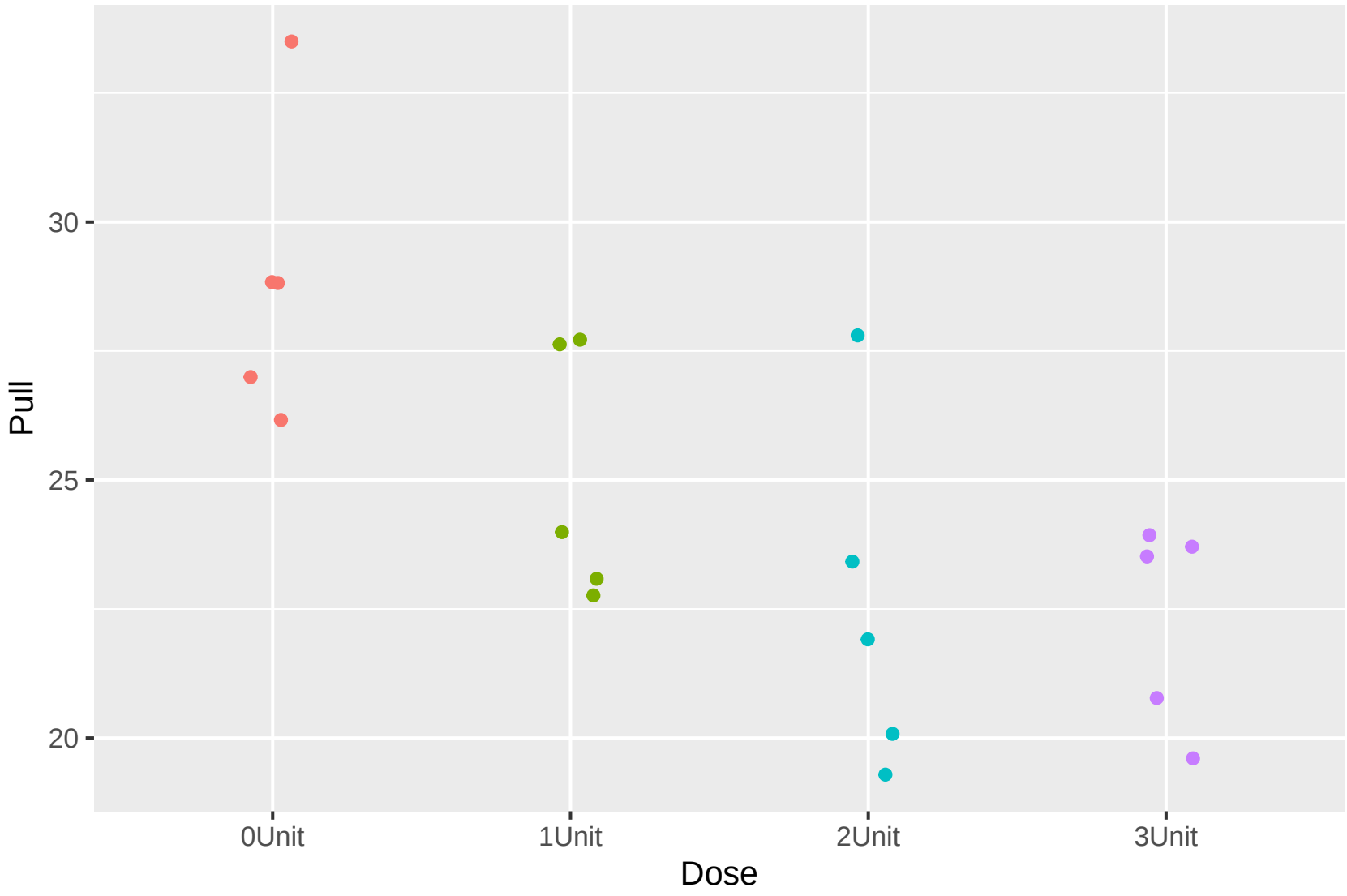
```
> tapply(Pull$Pull,Pull$Dose,mean)  #Group Means

+ 0Unit 1Unit 2Unit 3Unit
+ 28.86 25.04 22.50 22.30

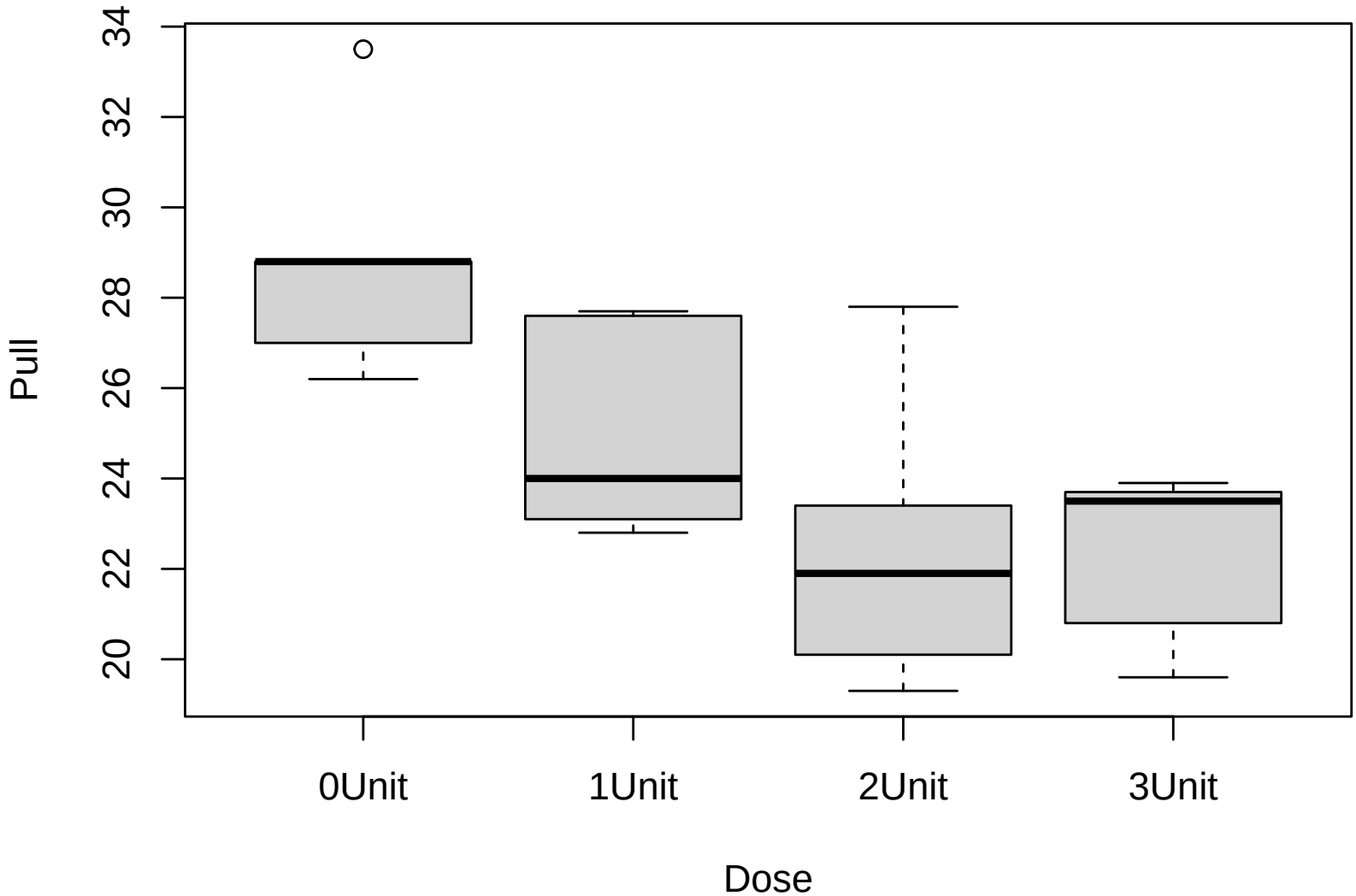
> tapply(Pull$Pull,Pull$Dose,var)   #Group Variances

+  0Unit  1Unit  2Unit  3Unit
+  8.018  5.873 11.315  3.875
```

Example: Anxiety Response



Example: Anxiety Response



Example: Anxiety Response

```
> summary(aov(Pull~Dose,data=Pull))

+             Df Sum Sq Mean Sq F value  Pr(>F)
+ Dose         3  140.1   46.70   6.423 0.00462 **
+ Residuals   16  116.3    7.27
+ ---
+ Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

> qf(0.95,3,16)

+ [1] 3.238872

> 1-pf(6.423,3,16)

+ [1] 0.004623413
```

Assessing unequal group variances

Here

	DOSE 0	DOSE 1	DOSE 2	DOSE 3
s_i^2	8.018	5.873	11.315	3.875

so we might suspect that the treatment variances $\sigma_1^2, \sigma_2^2, \sigma_3^2, \sigma_4^2$ are not equal. We may test this formally using

Levene's test.

Assessing unequal group variances

Note: Visual inspection can give an idea of whether the equal variance assumption is valid, or whether the populations are normal.

However, the sample sizes may be small, so that visual inspection or testing may not detect deviations from these assumptions.

Ideally we would like to be able to relax these assumptions.

Assessing unequal group variances

Levene's Test: To test

$$\begin{aligned} H_0 &: \sigma_1^2 = \sigma_2^2 = \cdots = \sigma_k^2 \\ H_1 &: \text{At least one pair of } \sigma^2 \text{ different.} \end{aligned}$$

Test statistic

$$W = \frac{(n-k)}{(k-1)} \frac{SST_Z}{SSE_Z} = \frac{MST_Z}{MSE_Z}$$

where SST_Z and SSE_Z are the usual sums of squares evaluated for the new data z_{ij} where now

$$z_{ij} = \left| x_{ij} - \overline{x}_i \right|.$$

If H_0 is true

$$W \sim F_{k-1,n-k}.$$

Assessing unequal group variances

```
> library(car)

+ Loading required package:  carData

> leveneTest(Pull~Dose,center='mean',data=Pull)

+ Levene's Test for Homogeneity of Variance (center = "mean")
+       Df F value Pr(>F)
+ group  3  0.2948 0.8286
+       16

> leveneTest(Pull~Dose,center='median',data=Pull)

+ Levene's Test for Homogeneity of Variance (center = "median")
+       Df F value Pr(>F)
+ group  3  0.1955 0.8979
+       16
```

The ANOVA Table

For a completely randomized design, we may report the results of the ANOVA F-test in a stylized form.

SOURCE	DF	SS	MS	F
TREATMENTS	$k-1$	SST	$MST = \frac{SST}{(k-1)}$	$F = \frac{MST}{MSE}$
ERROR	$n-k$	SSE	$MSE = \frac{SSE}{(n-k)}$	
TOTAL	$n-1$	SS		

The ANOVA Table

Note

(i) $(k-1)+(n-k)=(n-1)$

(ii) $SST + SSE = SS$

i.e. we can fill in missing values if they are not given.

The ANOVA Table

Sometimes an extra column is added at the right of the table to give the p-value of the ANOVA F-test.

SOURCE	DF	SS	MS	F	p
TMT	$k-1$	SST	MST	$F = \frac{MST}{MSE}$	p
ERROR	$n-k$	SSE	MSE		
TOTAL	$n-1$	SS			

where p solves

$$\frac{MST}{MSE} = F_{p,k-1,n-k}$$

and F_{a,ν_1,ν_2} is the upper a point of the $F_{k-1,n-k}$ distribution.

The ANOVA Table

```
> supernova(lm(Pull~Dose,data=Pull))

+  Analysis of Variance Table (Type III SS)
+  Model: Pull ~ Dose
+
+                              SS df     MS     F   PRE     p
+  ----- --------------- | ------- -- ------ ----- ----- -----
+  Model (error reduced) | 140.093  3 46.698 6.423 .5463 .0046
+  Error (from model)    | 116.324 16  7.270
+  ----- --------------- | ------- -- ------ ----- ----- -----
+  Total (empty model)   | 256.417 19 13.496
```

PRE is the proportional reduction in error

$$PRE = \frac{SST}{SS}$$

Further Examples and R code

- Milk: milk-protein level example
- Pull: pull-strength in Alzheimer's example
- AIDS:
  - RESPONSE: gut permeability of drug mannitol
  - FACTOR: AIDS HIV Status
  - FACTOR LEVELS: $k = 4$

    AIDS - Full AIDS

    ARC - AIDS-related conditions

    HIV+ - HIV positive

    HIV- - HIV negative

Further Examples and R code

- MilkBac: bacteria level in different batches of milk
  - RESPONSE: Bacteria level count per million
  - FACTOR: Batch number
  - FACTOR LEVELS: 1,2,3,4,5 ($k = 5$)

Further Examples and R code

- PTSD: PTSD score in different treatment groups.
  - RESPONSE: PTSD score
  - FACTOR: Therapeutic treatment method
  - FACTOR LEVELS: $k = 4$

    SIT - "Stress Innoculation Therapy"

    RE - "Relive Experience"

    SC - "Standard Counselling"

    WL - "Waiting List" (Control)

Further Examples and R code

Example: PTSD Example

$n = 45, k = 4$.

F-statistic	$F_{obs} = 3.046$
Critical Value	$F_{0.05,3,41} \simeq 2.84$
	$F_{0.025,3,41} \simeq 3.46$
	$F_{0.01,3,41} \simeq 4.31$

Textbook tables give $F_{\alpha,3,40}$.

$$\Longrightarrow \text{Reject } H_0 \text{ at } \alpha = 0.05 \quad (p = 0.039).$$

but Levene's Test suggests that the assumption of equal variances is not valid.

Relaxing the assumptions: permutation tests

Why do we need the three assumptions ?

- independence
- Normality
- equal variances

- so that we can predict (under H_0) that

$$F \sim F_{k-1,n-k}$$

and complete the test (compute p-values and the rejection region).

Relaxing the assumptions: permutation tests

But our null hypothesis of interest is (essentially)

$$H_0 \ : \ \text{No difference between treatments}$$

Under this hypothesis, the treatment labels

should not matter

that is, we should be able to exchange the group labels, and not notice any major difference in the test statistic.

Relaxing the assumptions: permutation tests

This leads us to consider permutation or randomization tests.

i.e. we compute the test statistic for all possible relabellings consistent with H_0, retaining the group sample sizes, and use these values to compute the rejection region.

Suppose that there are N possible relabellings that give rise to test statistics

$$F_1, F_2, \ldots, F_N$$

Relaxing the assumptions: permutation tests

Then the rejection region for significance level α is the interval to the right of

$$N(1-\alpha)\text{th largest of the values } F_1, F_2, \ldots, F_N$$

and the p-value is

$$\frac{\text{Number of } F_1, F_2, \ldots, F_N \geqslant F_{obs}}{N}$$

where the observed test statistic is

$$F_{obs} = \frac{MST}{MSE}.$$

Relaxing the assumptions: permutation tests

Note: If the group sample sizes are $n_1, n_2, \ldots, n_k$ then

$$N = \frac{n!}{n_1! n_2! \ldots n_k!}$$

where

$$n! = n(n-1)(n-2)\ldots 3.2.1$$

("n factorial") - potentially very large.

Relaxing the assumptions: permutation tests

Example: PTSD Example

$$k = 4, n = 45 \qquad (n_1 = 14, n_2 = 10, n_3 = 11, n_4 = 10)$$

There are

$$\frac{45!}{14!10!11!10!} = 2.610 \times 10^{24}$$

possible relabellings: a very big number. We must compute

$$F = \frac{MST}{MSE}$$

for each relabelling. Here, $F_{obs} = 3.046$.

Relaxing the assumptions: permutation tests

Example: PTSD Example

Instead of computing all of the F values for this large number of permutations, we may approximate the distribution using Monte Carlo procedures

- we sample $N = 10000$ (say) data sets where the labels are permuted
- we approximate the null distribution using this sample

Relaxing the assumptions: permutation tests

Example: PTSD Example

Using the exact permutation approach, we compute for $\alpha = 0.05$

$$\begin{aligned} \text{CRITICAL VALUE} &: C_R = 2.844 \\ \text{p-VALUE} &: p = 0.040 \end{aligned}$$

Compare this with the ANOVA F-test values

$$\begin{aligned} \text{CRITICAL VALUE} &: C_R = 2.833 \\ \text{p-VALUE} &: p = 0.039 \end{aligned}$$

using the $F_{3,41}$ distribution.

Relaxing the assumptions: permutation tests

Example: PTSD Example

Thus we obtain virtually identical results; however

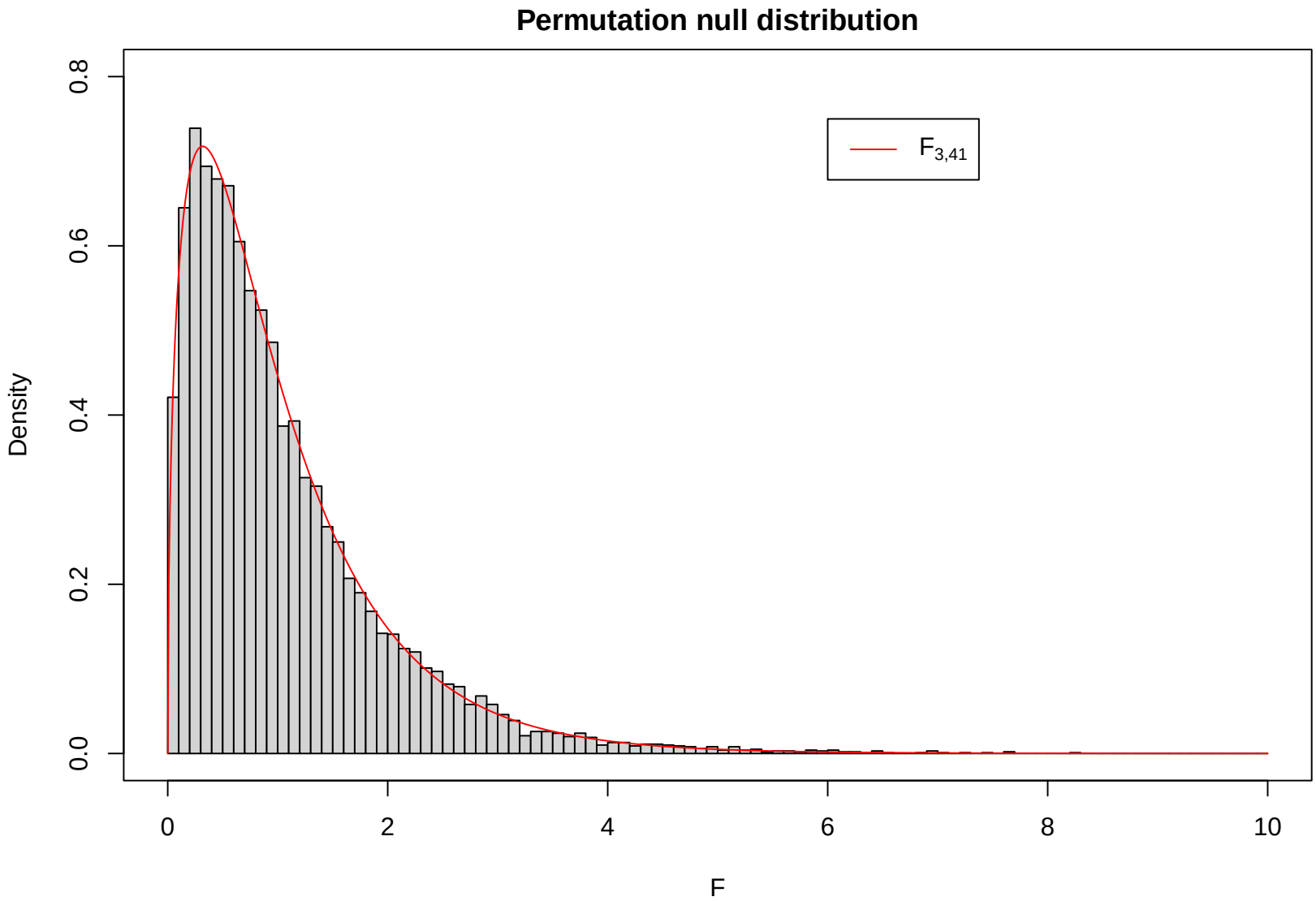
- the randomization test does not need the assumptions of Normality or equal variances.

Relaxing the assumptions: permutation tests

```
> N<-10000
> Fstat<-rep(0,N)
> n<-nrow(PTSD)
> x<-PTSD$PTSD
> z<-PTSD$GROUP
> for(iperm in 1:N){
+     gp<-z[sample(1:nrow(PTSD))]
+     Fstat[iperm]<-summary(aov(x~gp))[[1]][1,4]
+ }
> quantile(Fstat,prob=0.95)

+      95%
+ 2.846929
```

Relaxing the assumptions: permutation tests



Relaxing the assumptions: permutation tests

Example: PTSD Example

Thus the null hypothesis (of equal means) is

rejected

under both procedures at the $\alpha = 0.05$ significance level.

In this case, the computations give similar conclusions. Here the truth or otherwise of the normality/equal variance assumptions does not matter.

Final Note on ANOVA F-test for a CRD

If $k = 2$, consider $F = MST/MSE$;

$$\begin{aligned} MST &= \frac{1}{k-1}\sum_{i=1}^{k} n_i(\overline{x}_i - \overline{x})^2 = n_1(\overline{x}_1 - \overline{x})^2 + n_2(\overline{x}_2 - \overline{x})^2 \\ &= \frac{n_1 n_2}{n_1 + n_2}(\overline{x}_1 - \overline{x}_2)^2 \end{aligned}$$

$$\begin{aligned} MSE &= \frac{1}{n-k}\sum_{i=1}^{k}\sum_{j=1}^{n_i}(x_{ij} - \overline{x}_i)^2 = s_P^2 \\ &= \frac{(n_1-1)s_1^2 + (n_2-1)s_2^2}{n_1 + n_2 - 2} \end{aligned}$$

Final Note on ANOVA F-test for a CRD

Therefore

$$F_{obs} = \frac{\left(\frac{n_1 n_2}{n_1 + n_2}\right)(\overline{x}_1 - \overline{x}_2)^2}{s_P^2} = \left(\frac{(\overline{x}_1 - \overline{x}_2)}{s_p\sqrt{\frac{1}{n_1} + \frac{1}{n_2}}}\right)^2$$

Thus $F_{obs} = T_{obs}^2$, where T_{obs} is the two-sample t-test statistic.

Final Note on ANOVA F-test for a CRD

Thus if $k = 2$, the ANOVA F-test and the two sample t-test are equivalent

$$\begin{aligned} T &\sim t_{n-2} \\ F &\sim F_{1,n-2} \end{aligned}$$

and we must get the same conclusion (to reject H_0 or otherwise) using either statistic.

Summary

If the assumptions

- independence (holds by design in a CRD)
- Normal populations
- equal variances

hold, use

ANOVA F-test

If the assumptions do not hold

- use Randomization/Permutation test
- use Non-parametric test (see Section 3)

Multiple Comparison of Means

If the ANOVA F-test null hypothesis

$$H_0 : \mu_1 = \cdots = \mu_k$$

is rejected, then it is of interest to discover which of the means are different. For k groups, there are $c = k(k-1)/2$ pairs of group means that can be compared.

Multiple Comparison of Means

Consider a “family" of hypothesis tests - a collection of tests of different hypotheses carried out independently on different data sets. For each test in the family, we consider testing the hypothesis at significance level α.

Label the tests $i = 1, \ldots, c$, and for each i, label

- the null hypotheses H_{0i}
- the test statistics T_i
- the rejection regions $\mathcal{R}_i$

that are potentially different for each i.

Multiple Comparison of Means

We specify for each i,

$$\alpha = \Pr(T_i \in \mathcal{R}_i | H_{0i} \text{ is true})$$

which implicitly defines $\mathcal{R}_i$. Note that α is the

"Test Type I Error Rate" or "Comparisonwise Error Rate"

Now consider the results of all tests in the family; what is the "Familywise" Type I error rate ?

Multiple Comparison of Means

Using the laws of probability

$$\Pr(T_i \in \mathcal{R}_i | H_{0i} \text{ is true}) = \alpha$$

means that

$$\Pr(T_i \notin \mathcal{R}_i | H_{0i} \text{ is true}) = 1 - \alpha$$

giving the probability that the test does not reject H_{0i}, if H_{0i} is in fact true, is $1 - \alpha$.

Multiple Comparison of Means

Now we consider all tests together;

$$\Pr(\text{Each } T_i \notin \mathcal{R}_i | \text{Each } H_{0i} \text{ is true}) = (1-\alpha)^c$$

This is the probability that each test results in the null hypothesis not being rejected

- that is, the probability that we never commit a Type I error across the c tests.

Multiple Comparison of Means

Therefore the probability of at least one Type I error is

$$\alpha_F = 1 - (1 - \alpha)^c$$

α_F is the Familywise Error Rate.

	$\alpha = 0.05$	$\alpha = 0.01$
c	α_F	α_F
5	0.226	0.049
10	0.401	0.096
50	0.923	0.395
100	0.994	0.634

Therefore, when we carry out a “family" of tests, we should not use the traditional choices of $\alpha = 0.05$ or 0.01 on each test.

The Bonferroni Method

To fix $\alpha_F = 0.05$, say, we need to use α on each test where

$$\alpha_F = 1 - (1 - \alpha)^c \quad \Longleftrightarrow \quad \alpha = 1 - (1 - \alpha_F)^{1/c}$$

For example, if $\alpha_F = 0.05$ and $c = 10$, use

$$\alpha = 1 - (1 - 0.05)^{1/10} = 0.0051$$

The Bonferroni Method

It can be shown that

$$1 - (1 - \alpha)^c \approx c\alpha.$$

If α_F is the required familywise error rate, we must set the comparisonwise error rate to be $\alpha = \alpha_F/c$.

α_F/c is known as the Bonferroni Correction.

Confidence Intervals

For the $k = 2$ comparison of means, a $100(1-\alpha)\%$ confidence interval for $\mu_1 - \mu_2$ is

$$(\overline{x}_1 - \overline{x}_2) \pm t_{\alpha/2, n_1+n_2-2} s_P \sqrt{\frac{1}{n_1} + \frac{1}{n_2}}$$

where $t_{\alpha,\nu}$ is the upper α point of the t-distribution with ν degrees of freedom (under the assumptions of independence, Normality and equal group variances).

Confidence Intervals

Note: If we move to a family of c tests, to get simultaneous confidence intervals for the differences in means $\mu_i - \mu_j$ for all pairs of i and j, we should adjust α to α_F when computing the $100(1-\alpha)\%$ confidence interval.

Confidence Intervals

There are many different methods for correcting the confidence interval for use in different experimental situations.

For example, we might consider

- planned comparisons $\mu_1 = \mu_3$, $\mu_7 = \mu_{10}$ etc.
- all comparisons

of group means.

Confidence Intervals

Three methods are recommended:

- Bonferroni's Method
- Tukey's HSD Method for all pairwise comparisons
- Scheffé's Method for all possible contrast comparisons.

Having selected a multiple comparison correction method, we compute simultaneous confidence intervals for each comparison of means, and identify

- which means are significantly different
- the ranking of differences $\mu_i - \mu_j$ in terms of magnitude.

Randomized Block Designs

A randomized block design used matched experimental units organized into sets known as blocks and assigns one member from the set to each treatment.

For k treatments

1. Compile b blocks of k experimental units, with each block comprising units that are similar.
2. Assign one unit from each block to each treatment at random.

Then there are a total of $n = bk$ measured responses.

Randomized Block Designs

We wish to compare treatments whilst acknowledging that there may be differences between the blocks.

That is, the observed variation is due to

treatments and blocks and error

rather than merely

treatments and error

as was the case for the CRD.

Randomized Block Designs

Example: Test scores

- Response : Measured test Score
- Factor : Test type
- Factor-levels : $k = 2$ (Type 1, Type 2)
- Blocks : $b = 5$ (Previous GPA)

i.e. $k = 2, b = 5$, so $n = 10$.

Randomized Block Designs

Example: Test scores

Block	GPA	Type 1	Type 2
1	2.75	540	530
2	3.00	570	550
3	3.25	590	580
4	3.50	640	620
5	3.75	690	690

Randomized Block Designs

Example: Test scores

This design recognizes that GPA score is likely to explain some variation in test score, but is not directly related to the factor of interest (test type).

i.e. the blocking variable removes systematic variation in response that is not of primary interest.

Randomized Block Designs

Example: Test Scores

To implement, we

- pick two students from each GPA band (block)
- randomly assign one to take test type 1, the other to test type 2.

Randomized Block Designs

Example: Treatment for Hypertension

- Response : Blood Pressure (mgHg)
- Factor : Drug Type
- Factor-levels : $k = 3$ (Drug 1, Drug 2, Drug 3)
- Blocks : $b = 4$ Age/Sex combinations
  - Female/Under 50
  - Male/Under 50
  - Female/Over 50
  - Male/Over 50

i.e. $k = 3, b = 4 \therefore n = 12$.

Testing for Equal Treatment Means

An ANOVA F-test can be constructed for a RBD. Let

- $i = 1, \ldots, k$ index treatments
- $j = 1, \ldots, b$ index blocks

i.e. x_{ij} is the response for the ith treatment in the jth block. Let

- $\overline{x}_i$ be the ith treatment mean
- $\overline{x}_j^{(\mathrm{B})}$ be the jth block mean
- $\overline{x}$ be the overall mean

Testing for Equal Treatment Means

Let

$$
\begin{aligned}
SST &= \sum_{i=1}^{k} b\left(\overline{x}_i - \overline{x}\right)^2 \\
SSB &= \sum_{j=1}^{b} k\left(\overline{x}_j^{(B)} - \overline{x}\right)^2 \\
SS &= \sum_{i=1}^{k}\sum_{j=1}^{b}\left(x_{ij} - \overline{x}\right)^2
\end{aligned}
$$

SST: Sum of Squares for Treatments

SSB: Sum of Squares for Blocks

SS: Total Sum of Squares

Testing for Equal Treatment Means

Finally

$$SS = SST + SSB + SSE \quad \therefore \quad SSE = SS - SST - SSB$$

SSE: Sum of Squares for Errors

Test statistic is

$$F = \frac{MST}{MSE}$$

where

$$MST = \frac{SST}{k-1} \qquad MSE = \frac{SSE}{n-b-k+1}$$

Testing for Equal Treatment Means

The presumed model is

$$X_{ij} = \mu_i + \mu_j^{(\mathrm{B})} + \epsilon_{ij}$$

where

- μ_i is the mean contribution from treatment level i
- $\mu_j^{(\mathrm{B})}$ is the mean contribution from block j

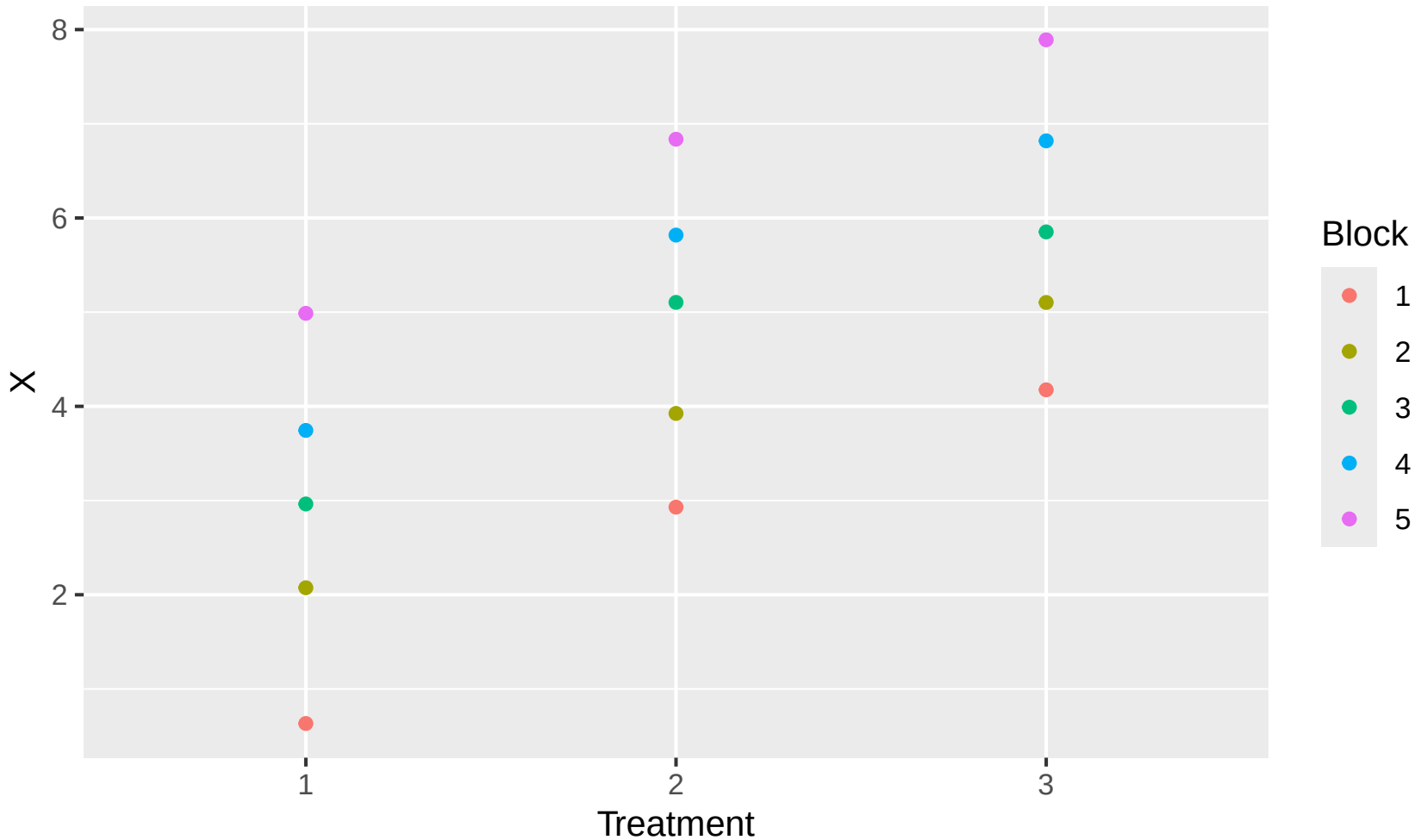
and the effects combine additively.

Testing for Equal Treatment Means

Here μ_i is not simply the population mean of the ith treatment group:

- the expected response for any experimental unit depends on the block they are in
- μ_i is the contribution to this expected response that is due to being assigned to treatment level i.

Testing for Equal Treatment Means



ANOVA F-test for a randomized block design

Theorem: ANOVA F-test for a RBD

To test

$$\begin{aligned} H_0 &: \mu_1 = \cdots = \mu_k \\ H_a &: \text{At least one pair of } \mu_i\text{s different.} \end{aligned}$$

use the test statistic

$$F = \frac{MST}{MSE}$$

If H_0 is true

$$F \sim F_{k-1,n-b-k+1}$$

This defines the rejection region for significance level α, and the p-value, in the usual way.

ANOVA F-test for a randomized block design

Assumptions:

1. Experimental units are independent, and treatments are allocated at random (within blocks).
2. Normality
3. bk block/treatment combinations correspond to populations with equal variances.

ANOVA F-test for a randomized block design

ANOVA Table

SOURCE	DF	SS	MS	F
TMTS	$k-1$	SST	MST	F = MST/MSE
BLOCKS	$b-1$	SSB	MSB	
ERROR	$n-k-b+1$	SSE	MSE	
TOTAL	$n-1$	SS		

ANOVA F-test for a randomized block design

After the ANOVA test is complete, and the hypothesis

$$H_0 : \mu_1 = \cdots = \mu_k$$

is rejected, we can proceed with the “post-hoc" tests of hypotheses $\mu_i = \mu_j$ for $i \neq j$.

ANOVA F-test for a randomized block design

Notes:

1. In a RBD, it is not (in general) possible to estimate individual treatment means: $\overline{x}_i$ does not estimate μ_i as it is an average across blocks, which are believed to be different in terms of response.

ANOVA F-test for a randomized block design

2. Testing the Block Means However, we can test whether the block means $\mu_1^{(B)}, \ldots, \mu_b^{(B)}$ are significantly different. For

$$H_0 : \mu_1^{(B)} = \cdots = \mu_b^{(B)}$$

we use the F statistic

$$F = \frac{MSB}{MSE}$$

where

$$MSB = \frac{SSB}{b-1}$$

ANOVA F-test for a randomized block design

If H_0 is true

$$F \sim F_{b-1,n-k-b+1}$$

That is, we treat the blocks as levels of another factor, and test to see whether this factor affects response.

ANOVA F-test for a randomized block design

Example: Soil Analysis (see handout)

Results of two ANOVA F-tests:

Test of	F	p	Conclusion
SOLVENT	0.673	0.585	No Difference
SOIL	10.568	0.001	Difference

Here SOLVENT is the treatment variable, SOIL is the blocking variable.

ANOVA F-test for a randomized block design

We must remember to check the assumptions (independence, normality, equal variances in each treatment/block combination)

Equal variances may be hard to check as we only have one observation per treatment/block comparison.

ANOVA F-test for a randomized block design

In the sums-of-squares decompositions

$$CRD: \quad SS = [SST] + SSE$$

$$RBD: \quad SS = [SST + SSB] + SSE$$

are both of the form

$$\begin{array}{ccccc} \text{TOTAL} & = & \text{SYSTEMATIC} & + & \text{RANDOM} \\ \text{VARIATION} & & \text{VARIATION} & & \text{VARIATION} \end{array}$$

ANOVA F-test for a randomized block design

"SYSTEMATIC" $\begin{cases} \text{For the CRD:} & SST \\ \text{For the RBD:} & SST + SSB \end{cases}$

"RANDOM" For both: SSE

ANOVA F-test for a randomized block design

We have studied the

Randomized Complete Block Design

where each block/treatment combination has one experimental unit.

An incomplete design could also be considered, where some block/treatment combinations are omitted. However, this design does not lead to straightforward ANOVA analysis.

Factorial Experiments

Designs studied so far:

- CRD - one factor
- RBD - one factor, plus one blocking variable, so two factors in total, where one (the blocking variable) is a known source of systematic variation.

However, in the RBD, we must assume that the treatments behave in a similar way across blocks.

Factorial Experiments

Let i index treatments ($1 \leqslant i \leqslant k$) and consider block j, and two treatment (factor levels) i_1 and i_2.

In an RBD, we assume that

$$\mathbb{E}[X_{i_1 j} - X_{i_2 j}] = \mu_{i_1} - \mu_{i_2}$$

which does not depend on j.

That is, the expected difference in response due to the two treatments does not depend on the block.

Factorial Experiments

But perhaps the difference does depend on block; perhaps we have interaction:

- an interaction is present if the effect of one variable on the outcome is modified in the presence of a second variable
- for example, the effect of treatment may be different in different blocks.

In the current RBD, we do not have enough data to look for this. We now seek to extend the RBD to allow for tests for interaction; we do this by using replication.

RBD with balanced replication

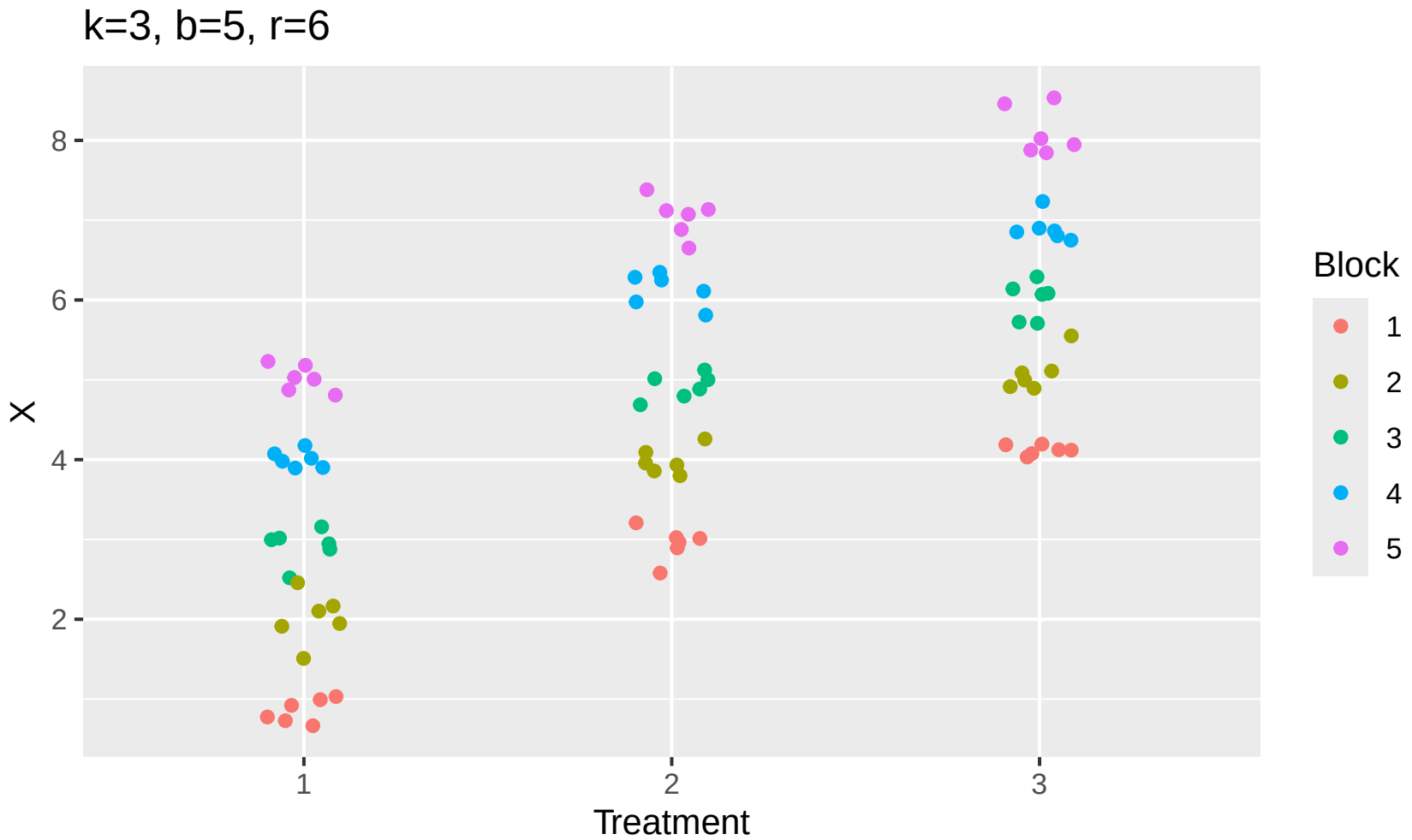
Suppose we have r observations per block/treatment combination (termed replicates), so that we have

$$n = bkr$$

experimental units in total.

Balanced designs have equal numbers of replicates in each block/treatment combination.

RBD with balanced replication



RBD with balanced replication

In this design, all the quantities

$$SST, SSB, SSE, SS$$

$$MST, MSB, MSE$$

can be defined, and an ANOVA F-test can be carried out - the only difference is that $n = bkr$.

RBD with balanced replication

- Sum of Squares for Treatments (SST)

$$SST = \sum_{i=1}^{k} br(\overline{x}_i - \overline{x})^2$$

- Sum of Squares for Blocks (SSB)

$$SSB = \sum_{j=1}^{b} kr(\overline{x}_j^{(B)} - \overline{x})^2$$

- Overall Sum of Squares (SS)

$$SS = \sum_{i=1}^{k} \sum_{j=1}^{b} \sum_{t=1}^{r} (x_{ijt} - \overline{x})^2$$

and $SSE = SS - SST - SSB$. The third index t indexes the replicates.

RBD with balanced replication

The RBD with replication does allow the investigation of interaction. The new test is based on the decomposition

$$SS = SST + SSB + SSI + SSE$$

where SSI is the Sum of Squares for Interaction.

RBD with balanced replication

We have SST, SSB and SS as before, and

$$SSI = \sum_{i=1}^{k} \sum_{j=1}^{b} r\left(\overline{x}_{ij} - \overline{x}_{i} - \overline{x}_{j}^{(B)} + \overline{x}\right)^2$$

where

$$\overline{x}_{ij} = \frac{1}{r} \sum_{t=1}^{r} x_{ijt} \qquad i = 1, \ldots, k,\ j = 1, \ldots, b$$

is the sample mean for replicates in (i, j)th treatment/block combination.

RBD with balanced replication

The presumed model here is

$$X_{ijt} = \mu_i + \mu_j^{(\mathrm{B})} + \mu_{ij}^{(\mathrm{I})} + \epsilon_{ijt}$$

where

- μ_i is the mean contribution from treatment level i
- $\mu_j^{(\mathrm{B})}$ is the mean contribution from block j
- $\mu_{ij}^{(\mathrm{I})}$ is the additional mean contribution for treatment i in block j

Note: we need to take care in interpreting these parameters.

Testing in the RBD with Replication

The three F statistics

$$F = \frac{MST}{MSE} \qquad F = \frac{MSB}{MSE} \qquad F = \frac{MSI}{MSE}$$

can be used to test for significant Treatment, Block and Interaction effects respectively.

Testing in the RBD with Replication

Now

$$MSE = \frac{SSE}{\text{Error d.f.}}$$

But what is “Error d.f." ? It is a constant that dictates how large SSE should be on average.

- the error d.f. depends on the complexity of the model fitted.

Testing in the RBD with Replication

The general rule for computing the error d.f. for any model is

$$\text{Error d.f.} = n - p$$

where n is the total sample size and p is the total number of parameters fitted.

Testing in the RBD with Replication

How many parameters do we fit ?

- No Interaction

$$p = 1 + (b - 1) + (k - 1)$$

that is, the overall mean μ, plus the $b-1$ differences from μ due to the blocks, plus the $k-1$ differences from μ due to the treatments.

- Interaction

$$p = bk$$

that is, one parameter in each cell of the two-way table of blocks by treatments.

Testing in the RBD with Replication

Thus

- No Interaction

$$p = 1 + (b-1) + (k-1) = b + k - 1$$

parameters, so

$$\text{Error d.f.} = n - p = n - b - k + 1$$

- Interaction: we fit $p = bk$ parameters, so

$$\text{Error d.f.} = n - p = n - bk$$

Testing in the RBD with Replication

It transpires that if

$$MSI = \frac{SSI}{(b-1)(k-1)}$$

is the Mean Square for Interaction, then

$$F = \frac{MSI}{MSE}$$

yields a test statistic suitable for testing interaction.

Testing in the RBD with Replication

Under the null hypothesis of no interaction, then

$$F \sim F_{(b-1)(k-1),n-bk}$$

where $n = bkr$.

Why $(b-1)(k-1)$? This is the number of extra parameters we fit to include the interaction.

Testing in the RBD with Replication

All of the ANOVA F-tests here are tests of model simplification

$$\begin{aligned} H_0 &: \text{NULL MODEL} \\ H_a &: \text{FULL MODEL} \end{aligned}$$

where we obtain the null model from the full model by setting some parameters to zero.

Testing in the RBD with Replication

For the CRD:

$$\begin{array}{rcl} H_a & & H_0 \\ \text{FULL MODEL} & \longrightarrow & \text{NULL MODEL} \\ k \text{ parameters} & \longrightarrow & 1 \text{ parameter} \end{array}$$

so there are $(k-1)$ extra parameters, and so SST varies on $(k-1)$ degrees of freedom

$$\mu_1, \ldots, \mu_k \longrightarrow \mu$$

Testing in the RBD with Replication

The (i,j)th treatment/block combination has mean

$$\mu_i + \mu_j^{(\mathrm{B})}$$

so for testing for a TREATMENT effect

$$\begin{array}{rcl} & H_a & H_0 \\ \text{FULL MODEL} & \longrightarrow & \text{NULL MODEL} \\ 1+(k-1)+(b-1)\text{ param.} & \longrightarrow & 1+(b-1)\text{ param.} \end{array}$$

so there are $(k-1)$ extra parameters, and so SST varies on $(k-1)$ degrees of freedom.

$$\mu_1, \ldots, \mu_k \longrightarrow \mu$$

Testing in the RBD with Replication

For testing for a BLOCK effect

$$\begin{array}{rcl} & H_a & H_0 \\ & \text{FULL MODEL} & \longrightarrow \quad \text{NULL MODEL} \\ 1+(k-1)+(b-1)\text{ param.} & \longrightarrow & 1+(k-1)\text{ param.} \end{array}$$

so there are $(b-1)$ extra parameters, and so SSB varies on $(b-1)$ degrees of freedom

$$\mu_1^{(\mathrm{B})}, \ldots, \mu_k^{(\mathrm{B})} \longrightarrow \mu^{(\mathrm{B})}$$

These models and tests can be fitted and carried out even if we do not have replication.

Testing in the RBD with Replication

With replication, we can investigate the interaction, that is the model where the (i,j)th treatment/block combination has mean

$$\mu_i + \mu_j^{(\mathrm{B})} + \mu_{ij}^{(\mathrm{I})}$$

rather than the model where

$$\mu_i + \mu_j^{(\mathrm{B})}$$

that is, we wish to test

$$\begin{aligned} H_0 &: \mu_{ij}^{(\mathrm{I})} = 0 \qquad \text{for all } i \text{ and } j \\ H_a &: \mu_{ij}^{(\mathrm{I})} \neq 0 \end{aligned}$$

Testing in the RBD with Replication

In the full interaction model: we fit bk parameters

In the restricted, no interaction model: we fit

$$1 + (b-1) + (k-1) = b + k - 1$$

parameters. Therefore the differences is

$$bk - (b + k - 1) = bk - b - k + 1 = (b-1)(k-1)$$

and SSI varies on $(b-1)(k-1)$ degrees of freedom.

Testing in the RBD with Replication

If we do not have replication, we CANNOT fit the interaction. Recall that

$$\text{Error d.f.} = n - bk$$

but if $r = 1$, $n = rbk = bk$, so the error d.f. is zero.

In fact, $SSE = 0$ also, so the MSE is not defined.

ANOVA Table

SOURCE	DF	SS	MS	F
TMTS	$k-1$	SST	MST	F_T
BLOCKS	$b-1$	SSB	MSB	F_B
INTERACTION	$(b-1)(k-1)$	SSI	MSI	F_I
ERROR	$(n-bk)$	SSE	MSE	
TOTAL	$n-1$	SS		

where

$$MST = \frac{SST}{k-1} \qquad MSB = \frac{SSB}{b-1}$$

$$MSI = \frac{SSI}{(b-1)(k-1)} \qquad MSE = \frac{SSE}{n-bk}$$

and

$$F_T = \frac{MST}{MSE} \qquad F_B = \frac{MSB}{MSE} \qquad F_I = \frac{MSI}{MSE}$$

Example

Simulated data with $k = 5, b = 3, r = 6$:

```
> k<-3
> b<-5
> r<-6
> tmt<-rep(rep(1:k,each=b),r)
> blk<-rep(rep(1:b,k),r)
> rept<-rep(1:r,b*k)
> mui<-c(0,2,3)
> mub<-c(1:5)
> X<-mui[tmt]+mub[blk]+rnorm(r*b*k)*0.2
> RBDdat<-data.frame(Treatment=as.factor(tmt),Block=as.factor(blk),X)
> supernova(lm(X~Treatment+Block+Treatment:Block,data=RBDdat), type=2)

+  Analysis of Variance Table (Type II SS)
+  Model: X ~ Treatment + Block + Treatment:Block
+
+                                      SS df     MS        F   PRE     p
+  -------------- -------------- | ------- -- ------ -------- ----- -----
+            Model (error reduced) | 321.762 14 22.983  533.074 .9901 .0000
+        Treatment                 | 136.912  2 68.456 1587.787 .9769 .0000
+            Block                 | 184.673  4 46.168 1070.838 .9828 .0000
+  Treatment:Block                 |   0.177  8  0.022    0.515 .0520 .8419
+            Error (from model)    |   3.234 75  0.043
+  -------------- -------------- | ------- -- ------ -------- ----- -----
+            Total (empty model)   | 324.995 89  3.652
```

Example

Interaction model:

```
> summary(aov(X~Treatment+Block+Treatment:Block,data=RBDdat))

+                 Df Sum Sq Mean Sq  F value Pr(>F)
+ Treatment        2 136.91   68.46 1587.787 <2e-16 ***
+ Block            4 184.67   46.17 1070.838 <2e-16 ***
+ Treatment:Block  8   0.18    0.02    0.515  0.842
+ Residuals       75   3.23    0.04
+ ---
+ Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

Example

No interaction model:

```
> summary(aov(X~Treatment+Block,data=RBDdat))

+             Df Sum Sq Mean Sq F value Pr(>F)
+ Treatment    2 136.91   68.46    1666 <2e-16 ***
+ Block        4 184.67   46.17    1123 <2e-16 ***
+ Residuals   83   3.41    0.04
+ ---
+ Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

Example: Batteries Data (see handout)

```
> summary(aov(life~temp+material+temp:material, data=Batteries))

+               Df Sum Sq Mean Sq F value      Pr(>F)
+ temp           2  39083   19542  29.344 0.000000169 ***
+ material       2  10633    5317   7.983     0.00189 **
+ temp:material  4   9438    2359   3.543     0.01897 *
+ Residuals     27  17981     666
+ ---
+ Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

For $\alpha = 0.05$, there is a significant temp effect ($p < 0.001$), and a significant material effect ($p < 0.002$), and a significant interaction ($p < 0.019$)

Multifactor designs

We now study multifactor designs, to assess the effects and interactions of several factors simultaneously.

We consider all possible combinations of

$$\begin{array}{lll} \text{FACTOR A} & \text{with} & a \text{ levels} \\ \text{FACTOR B} & \text{with} & b \text{ levels} \\ \text{FACTOR C} & \text{with} & c \text{ levels} \\ & \vdots & \end{array}$$

to define the treatments in a factorial design.

Factorial Experiments

A complete factorial experiment is one in which every combination of the factors is utilized.

- the number of treatments is equal to the total number of factor-level combinations.

Consider two factor experiments:

$$\begin{array}{lll} \text{FACTOR A} & \text{with} & a \text{ levels} \\ \text{FACTOR B} & \text{with} & b \text{ levels} \end{array}$$

so there are ab treatments in total.

Factorial Experiments

The data can be arranged in a table where the cells define the factor-level combinations.

		Factor B				
		1	2	3	4	5
Factor A	1					
	2					
	3					

This table displays a two-way layout with $a = 3$ and $b = 5$. Here

- Levels of A will be in the rows
- Levels of B will be in the columns

Factorial Experiments

This design is very similar to the RBD, but now the second factor is not a blocking factor:

- the ab treatment populations are constructed independently from the same base population, and are believed to be not systematically different.
- individuals from the same base population are assigned at random to one of the ab treatments.

Factorial Experiments

In this design we can study the effect of

- Factor A and Factor B (the main effects)
- the interaction

provided we have balanced replication.

Factorial Experiments

We construct ANOVA F-tests based on the decomposition

$$SS = SST_A + SST_B + SSI_{AB} + SSE$$

- Sum of Squares for Treatments due to factor A (SST_A)

$$SST_A = \sum_{i=1}^{a} br(\overline{x}_{i.} - \overline{x}_{..})^2$$

- Sum of Squares for Treatments due to factor B (SST_B)

$$SST_B = \sum_{j=1}^{b} ar(\overline{x}_{.j} - \overline{x}_{..})^2$$

- Sum of Squares for Interaction (SSI_AB)

$$SSI_{AB} = \sum_{i=1}^{a}\sum_{j=1}^{b} r(\overline{x}_{ij} - \overline{x}_{i.} - \overline{x}_{.j} + \overline{x}_{..})^2$$

Factorial Experiments

New notation:

- sample mean for Factor A level i

$$\overline{x}_{i.} = \frac{1}{br} \sum_{j=1}^{b} \sum_{t=1}^{r} x_{ijt} \qquad i = 1, \ldots, a$$

  which is a row mean

- sample mean for Factor B level j

$$\overline{x}_{.j} = \frac{1}{ar} \sum_{i=1}^{a} \sum_{t=1}^{r} x_{ijt} \qquad j = 1, \ldots, b$$

  which is a column mean

Factorial Experiments

- sample mean for replicates in (i,j)th factor combination

$$\overline{x}_{ij} = \frac{1}{r}\sum_{t=1}^{r} x_{ijt}$$

  for each $i = 1,\ldots,a,\ j = 1,\ldots,b,$ which is a cell mean

- overall sample mean

$$\overline{x}_{..} = \frac{1}{n}\sum_{i=1}^{a}\sum_{j=1}^{b}\sum_{t=1}^{r} x_{ijt}$$

  which is termed the overall mean

Factorial Experiments

These allow computation of

$$SST_A, SST_B, SSI_{AB}, SS, SSE$$

$$MST_A, MST_B, MSI_{AB}, MSE$$

using the degrees of freedom identical to those in the RBD with replication.

Tests for

- significant effect for Factor A
- significant effect for Factor B
- significant interaction

will be carried out as before using an ANOVA table.

ANOVA Table

SOURCE	DF	SS	MS	F
FACTOR A	$a-1$	SST_A	MST_A	F_A
FACTOR B	$b-1$	SST_B	MST_B	F_B
INTERACTION	$(a-1)(b-1)$	SSI_{AB}	MSI_{AB}	F_{AB}
ERROR	$(n-ab)$	SSE	MSE	
TOTAL	$n-1$	SS		

ANOVA Table

If Factor A is not influential (H_0 specifying no difference between responses at different levels of factor A), then

$$F_A \sim F_{a-1,n-ab}$$

Similarly,

$$\begin{aligned} \text{No effect of Factor B} &: F_B \sim F_{b-1,n-ab} \\ \text{No Interaction} &: F_{AB} \sim F_{(a-1)(b-1),n-ab} \end{aligned}$$

Notation

The main effects plus interaction model can be written

$$A + B + A{:}B$$

whereas the main effects only can be written

$$A + B$$

The models

$$A + A{:}B \qquad\qquad B + A{:}B$$

do not make sense.

Notation

For a two factor design, the only models that should be considered and reported are

MODEL	FACTOR	INTERACTION
NULL	NONE	NO
A	A	NO
B	B	NO
A+B	A,B	NO
A+B+A:B	A,B	YES

that is, if the interaction is significant, the only model you should report is

$$A + B + A{:}B$$

Two variable analysis

ANOVA analysis for the RBD and FD (both with replication) are identical. The only difference lies in the interpretation of the factors

- RBD: one blocking, one treatment factor
- FD: two treatment factors

“Blocking” factors are known or strongly believed to have a significant effect on the response.

Estimating Effect Size

In multifactor designs, parameter estimation can be carried out in different parameterizations

For the CRD (one-way layout):

- Natural parameters: $\mu_1, \ldots, \mu_k$
- Contrast parameters: $\beta_1, \beta_2, \ldots, \beta_k$ where

$$\begin{aligned} \beta_1 &= \mu_1 \\ \beta_i &= \mu_i - \mu_1, \quad i = 2, \ldots, k \end{aligned}$$

Estimating Effect Size

The $(k-1)$ parameters

$$\beta_2, \ldots, \beta_k$$

measure the differences in mean from baseline group (level 1 of the factor).

The null hypothesis assumes that

$$\beta_2 = \cdots = \beta_k = 0.$$

Estimating Effect Size

For the Milk protein data set: using the natural parameters

```
> summary(lm(protein~-1+Diet,data=Milk))$coef

+                   Estimate Std. Error  t value Pr(>|t|)
+ Dietbarley        3.531929 0.01551264 227.6807        0
+ Dietbarley+lupins 3.429695 0.01492705 229.7638        0
+ Dietlupins        3.312384 0.01502558 220.4497        0

> tapply(Milk$protein,Milk$Diet,mean)

+        barley barley+lupins        lupins
+      3.531929      3.429695      3.312384
```

Here the `Estimate` column contains the means for the three groups which are the estimates of

$$\mu_1, \mu_2, \mu_3$$

Estimating Effect Size

Using the contrast parameters:

```
> summary(lm(protein~Diet,data=Milk))$coef

+                     Estimate Std. Error    t value     Pr(>|t|)
+ (Intercept)        3.5319294 0.01551264 227.680728 0.000000e+00
+ Dietbarley+lupins -0.1022344 0.02152809  -4.748885 2.266185e-06
+ Dietlupins        -0.2195453 0.02159653 -10.165769 1.963302e-23
```

which reveals the estimates of β_2 and β_3 as

$$-0.1022344 \quad \text{and} \quad -0.2195453$$

Estimating Effect Size

For the two-factor designs (RBD/FD): The cell means can be denoted m_{ij} where

$$m_{ij} = \mu_{i.} + \mu_{.j} + \mu_{ij}$$

for cells (i,j), $i = 1,\ldots,a$, $j = 1,\ldots,b$, where

- $\mu_{i.}$ gives the Factor A contribution
- $\mu_{.j}$ gives the Factor B contribution
- μ_{ij} gives the interaction

and for the replicates in cell (i,j), we have

$$X_{ijt} \sim Normal(m_{ij}.\sigma^2).$$

Estimating Effect Size

By default, R uses the contrast parameterization:

$$
\begin{aligned}
m_{ij} &= \beta_1 && i = 1, j = 1 \\
&= \beta_1 + \beta_i^{(A)} && i = 2, \ldots, a, j = 1 \\
&= \beta_1 + \beta_j^{(B)} && i = 1, j = 2, \ldots, b \\
&= \beta_1 + \beta_i^{(A)} + \beta_j^{(B)} + \gamma_{ij}^{(AB)} && i = 2, \ldots, a \\
& && j = 2, \ldots, b
\end{aligned}
$$

where

$$
\begin{aligned}
\beta_i^{(A)} &: \text{contrasts for factor A} \\
\beta_j^{(B)} &: \text{contrasts for factor B} \\
\gamma_{ij}^{(AB)} &: \text{interaction}
\end{aligned}
$$

Estimating Effect Size

By default, R takes the first level of each factor as the baseline level, and looks at differences compared to this baseline.

The ab parameters are

β_1	1
$\beta_2^{(A)}, \ldots, \beta_a^{(A)}$	$(a-1)$
$\beta_2^{(B)}, \ldots, \beta_b^{(B)}$	$(b-1)$
$\gamma_{ij}^{(AB)}, i = 2, \ldots, a, j = 2, \ldots, b$	$(a-1)(b-1)$
Total	ab

Estimating Effect Size

For example: $a = 3, b = 4$.

Factor A \ Factor B	1	2	3	4
1	β_1	$\beta_1 + \beta_2^{(B)}$	$\beta_1 + \beta_3^{(B)}$	$\beta_1 + \beta_4^{(B)}$
2	$\beta_1 + \beta_2^{(A)}$	①	②	③
3	$\beta_1 + \beta_3^{(A)}$	④	⑤	⑥

where

$$① = \beta_1 + \beta_2^{(A)} + \beta_2^{(B)} + \gamma_{22}^{(AB)}$$

$$⑥ = \beta_1 + \beta_3^{(A)} + \beta_4^{(B)} + \gamma_{34}^{(AB)}$$

and so on.

Estimating Effect Size

Estimation is still straightforward:

PARAMETER	ESTIMATE
β_1	$\overline{x}_{11}$
$\beta_i^{(A)}$	$\overline{x}_{i1} - \overline{x}_{11}$
$\beta_j^{(B)}$	$\overline{x}_{1j} - \overline{x}_{11}$
$\gamma_{ij}^{(AB)}$	$\overline{x}_{ij} - \overline{x}_{i1} - \overline{x}_{1j} + \overline{x}_{11}$

for $i = 2, \ldots, a, j = 2, \ldots, b$.

Note: other contrast parameterizations can be used.

Example

For example: $a = 4, b = 3$.

- Factor A: levels $1, 2, \ldots, a$
- Factor B: levels $1, 2, \ldots, b$

Most complicated model: Main Effects plus Interaction

A + B + A:B

Example

That is, we have

- a baseline mean: β_1
- an effect for each level of Factor A: $\beta_i^{(A)}$
- an effect for each level of Factor B: $\beta_j^{(B)}$
- an interaction that modifies the effect of changing levels of Factor A at each level of Factor B: $\gamma_{ij}^{(AB)}$

Example

In R, the baseline group is the one where Factor A has level 1 and Factor B has level 1, but this choice is arbitrary

- changing this assumption should have no effect on the results we obtain.

Example

We adopt the following modelling strategy:

- Establish a baseline
- Look for changes from baseline introduced by Factor A
- Look for changes from baseline introduced by Factor B

Example

- Look for changes from baseline introduced by Factor A and Factor B acting additively, so that the effect of changing the level of Factor A is identical in each level of Factor B, and vice versa).

Example

- Look for changes from baseline introduced by Factor A and Factor B additively with interaction, so that the effect of changing the level of Factor A is different in each level of Factor B, and vice versa).

Two-way table: 4×3

Factor B

Factor A	1	2	3
1			
2			
3			
4			

Null Model : Baseline Mean Only

Null Model: no effect of either factor

Factor B

Factor A	1	2	3
1	β_1	β_1	β_1
2	β_1	β_1	β_1
3	β_1	β_1	β_1
4	β_1	β_1	β_1

Effect of Factor A only

Main Effect Only: A

Factor B

Factor A	1	2	3
1	β_1	β_1	β_1
2	$\beta_1 + \beta_2^{(A)}$	$\beta_1 + \beta_2^{(A)}$	$\beta_1 + \beta_2^{(A)}$
3	$\beta_1 + \beta_3^{(A)}$	$\beta_1 + \beta_3^{(A)}$	$\beta_1 + \beta_3^{(A)}$
4	$\beta_1 + \beta_4^{(A)}$	$\beta_1 + \beta_4^{(A)}$	$\beta_1 + \beta_4^{(A)}$

Effect of Factor B only

Main Effect Only: B

Factor B

Factor A	1	2	3
1	β_1	$\beta_1 + \beta_2^{(B)}$	$\beta_1 + \beta_3^{(B)}$
2	β_1	$\beta_1 + \beta_2^{(B)}$	$\beta_1 + \beta_3^{(B)}$
3	β_1	$\beta_1 + \beta_2^{(B)}$	$\beta_1 + \beta_3^{(B)}$
4	β_1	$\beta_1 + \beta_2^{(B)}$	$\beta_1 + \beta_3^{(B)}$

Effect of Factor A plus Effect of Factor B

Main Effects Only: A + B

Factor B

Factor A	1	2	3
1	β_1	$\beta_1 + \beta_2^{(B)}$	$\beta_1 + \beta_3^{(B)}$
2	$\beta_1 + \beta_2^{(A)}$	$\beta_1 + \beta_2^{(A)} + \beta_2^{(B)}$	$\beta_1 + \beta_2^{(A)} + \beta_3^{(B)}$
3	$\beta_1 + \beta_3^{(A)}$	$\beta_1 + \beta_3^{(A)} + \beta_2^{(B)}$	$\beta_1 + \beta_3^{(A)} + \beta_3^{(B)}$
4	$\beta_1 + \beta_4^{(A)}$	$\beta_1 + \beta_4^{(A)} + \beta_2^{(B)}$	$\beta_1 + \beta_4^{(A)} + \beta_3^{(B)}$

Main effects plus Interaction between A and B

Main Effects Plus Interaction: A + B + A:B

Factor B

Factor A	1	2	3
1	β_1	$\beta_1 + \beta_2^{(B)}$	$\beta_1 + \beta_3^{(B)}$
2	$\beta_1 + \beta_2^{(A)}$	$\beta_1 + \beta_2^{(A)} + \beta_2^{(B)} + \gamma_{22}^{(AB)}$	$\beta_1 + \beta_2^{(A)} + \beta_3^{(B)} + \gamma_{23}^{(AB)}$
3	$\beta_1 + \beta_3^{(A)}$	$\beta_1 + \beta_3^{(A)} + \beta_2^{(B)} + \gamma_{32}^{(AB)}$	$\beta_1 + \beta_3^{(A)} + \beta_3^{(B)} + \gamma_{33}^{(AB)}$
4	$\beta_1 + \beta_4^{(A)}$	$\beta_1 + \beta_4^{(A)} + \beta_2^{(B)} + \gamma_{42}^{(AB)}$	$\beta_1 + \beta_4^{(A)} + \beta_3^{(B)} + \gamma_{43}^{(AB)}$

Null Model : Baseline Mean Only

Null Model: no effect of either factor

Factor B

Factor A	1	2	3
1	10	10	10
2	10	10	10
3	10	10	10
4	10	10	10

Effect of Factor A only

Main Effect Only: A

Factor B

Factor A	1	2	3
1	10	10	10
2	10 + 2	10 + 2	10 + 2
3	10 + 5	10 + 5	10 + 5
4	10 + 8	10 + 8	10 + 8

Effect of Factor B only

Main Effect Only: B

Factor B

Factor A	1	2	3
1	10	10 + (-1)	10 + (-4)
2	10	10 + (-1)	10 + (-4)
3	10	10 + (-1)	10 + (-4)
4	10	10 + (-1)	10 + (-4)

Main effects of factors A and B only

Main Effects Only: A + B

Factor B

Factor A	1	2	3
1	10	10 + (-1)	10 + (-4)
2	10 + 2	10 + 2 + (-1)	10 + 2 + (-4)
3	10 + 5	10 + 5 + (-1)	10 + 5 + (-4)
4	10 + 8	10 + 8 + (-1)	10 + 8 + (-4)

Main effects plus Interaction between A and B

Main Effects Plus Interaction: A + B + A:B

Factor A	Factor B: 1	Factor B: 2	Factor B: 3
1	10	10 + (-1)	10 + (-4)
2	10 + 2	10 + 2 + (-1) + (-2)	10 + 2 + (-4) + 1
3	10 + 5	10 + 5 + (-1) + (-5)	10 + 5 + (-4) + 8
4	10 + 8	10 + 8 + (-1) + 2	10 + 8 + (-4) + 7

Illegal models

Q. Why are the following models

- A:B
- A + A:B
- B + A:B

not considered ?

A. Because they make specific, perhaps unrealistic assumptions about the data, and they imply that the levels of the factors are not arbitrarily labelled.

R will not fit such models, although it appears that it does !

Illegal models

Recall the definition of interaction:

- Variation in the effect of changing levels of one factor at the different levels of the other factor.
- For example, the effect on the response mean of moving from level 1 to level 2 for Factor B is different at different levels of Factor A.

Consider the model

A:B

this model implies that all parameters apart from the baseline and the interaction parameters are zero.

Illegal models

Interaction: A:B

Factor B

Factor A	1	2	3
1	β_1	$\beta_1 + 0$	$\beta_1 + 0$
2	$\beta_1 + 0$	$\beta_1 + 0 + 0 + \gamma_{22}^{(AB)}$	$\beta_1 + 0 + 0 + \gamma_{23}^{(AB)}$
3	$\beta_1 + 0$	$\beta_1 + 0 + 0 + \gamma_{32}^{(AB)}$	$\beta_1 + 0 + 0 + \gamma_{33}^{(AB)}$
4	$\beta_1 + 0$	$\beta_1 + 0 + 0 + \gamma_{42}^{(AB)}$	$\beta_1 + 0 + 0 + \gamma_{43}^{(AB)}$

Illegal models

In this set-up,

- for Factor A, Level 1: the effect of moving from Level 1 to Level 2 of factor B is zero
- for Factor A, Level 3: the effect of moving from Level 1 to Level 2 of factor B is $\gamma_{32}^{(AB)}$.

Therefore, there is a fundamental difference between the way that we regard the levels of Factor A.

Illegal models

Main Effects of Factor A Plus Interaction: A + A:B

Factor A \ Factor B	1	2	3
1	β_1	$\beta_1 + 0$	$\beta_1 + 0$
2	$\beta_1 + \beta_2^{(A)}$	$\beta_1 + \beta_2^{(A)} + 0 + \gamma_{22}^{(AB)}$	$\beta_1 + \beta_2^{(A)} + 0 + \gamma_{23}^{(AB)}$
3	$\beta_1 + \beta_3^{(A)}$	$\beta_1 + \beta_3^{(A)} + 0 + \gamma_{32}^{(AB)}$	$\beta_1 + \beta_3^{(A)} + 0 + \gamma_{33}^{(AB)}$
4	$\beta_1 + \beta_4^{(A)}$	$\beta_1 + \beta_4^{(A)} + 0 + \gamma_{42}^{(AB)}$	$\beta_1 + \beta_4^{(A)} + 0 + \gamma_{43}^{(AB)}$

Illegal models

In this set-up,

- for Factor A, Level 1: the effect of moving from Level 1 to Level 2 of Factor B is zero
- for Factor A, Level 3: the effect of moving from Level 1 to Level 2 of Factor B is $\gamma_{32}^{(AB)}$.

Illegal models

Therefore, there is a fundamental difference between the way that we regard the levels of Factor A. If we rearrange the labels of the levels of Factor A

we may get a different result.

Therefore, although it is possible in general to fit such models, it is no longer possible to talk of the effect of “Factor A".

How does R Handle Such Models ?

It is possible to use the models

A + A:B B + A:B A:B

in an ANOVA in R.

How does R Handle Such Models ?

$$A + B + A{:}B$$

```
> a<-3
> b<-5
> r<-6
> A<-rep(rep(1:a,each=b),r)
> B<-rep(rep(1:b,a),r)
> rept<-rep(1:r,b*a)
> mua<-c(0,2,3)
> mub<-c(1:5)
> X<-mua[A]+mub[B]+rnorm(r*b*a)*0.2
> FDdat<-data.frame(A=as.factor(A),B=as.factor(B),X)
> summary(aov(X~A+B+A:B, data=FDdat))

+             Df Sum Sq Mean Sq  F value Pr(>F)
+ A            2 141.50   70.75 1895.568 <2e-16 ***
+ B            4 181.50   45.37 1215.734 <2e-16 ***
+ A:B          8   0.37    0.05    1.229  0.294
+ Residuals   75   2.80    0.04
+ ---
+ Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

How does R Handle Such Models ?

A + A:B

```
> summary(aov(X~A+A:B, data=FDdat))

+             Df Sum Sq Mean Sq F value Pr(>F)
+ A            2  141.5   70.75  1895.6 <2e-16 ***
+ A:B         12  181.9   15.16   406.1 <2e-16 ***
+ Residuals   75    2.8    0.04
+ ---
+ Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

How does R Handle Such Models ?

This produces the usual ANOVA table, with terms including

A and A:B

However, in fact the model

A + B + A:B

has been fitted

- The results are just reported differently
- The terms B and A:B are reported together

How does R Handle Such Models ?

This is the default behaviour in the standard functions. However, it is not universal

- `supernova` exhibits different default behaviour
- `Anova` from the `car` library also has different behaviour

How does R Handle Such Models ?

```
> supernova(lm(X~A+B+A:B, data=FDdat))

+  Analysis of Variance Table (Type III SS)
+  Model: X ~ A + B + A:B
+
+                              SS df     MS       F   PRE     p
+  ----- --------------- | ------- -- ------ ------- ----- -----
+  Model (error reduced) | 323.364 14 23.097 618.850 .9914 .0000
+      A                 |  27.853  2 13.927 373.135 .9087 .0000
+      B                 |  57.798  4 14.449 387.143 .9538 .0000
+    A:B                 |   0.367  8  0.046   1.229 .1159 .2943
+  Error (from model)    |   2.799 75  0.037
+  ----- --------------- | ------- -- ------ ------- ----- -----
+  Total (empty model)   | 326.163 89  3.665

> summary(aov(X~A+B+A:B, data=FDdat))    #Previous analysis

+             Df Sum Sq Mean Sq  F value Pr(>F)
+ A            2 141.50   70.75 1895.568 <2e-16 ***
+ B            4 181.50   45.37 1215.734 <2e-16 ***
+ A:B          8   0.37    0.05    1.229  0.294
+ Residuals   75   2.80    0.04
+ ---
+ Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

How does R Handle Such Models ?

These results are different. To recover the original results, we need to specify that we want the

Type II sum of squares

which accounts for the presence of the interaction terms in a different way.

How does R Handle Such Models ?

```
> supernova(lm(X~A+B+A:B, data=FDdat), type=2)

+  Analysis of Variance Table (Type II SS)
+  Model: X ~ A + B + A:B
+
+                             SS df     MS        F   PRE     p
+  ----- --------------- | ------- -- ------ -------- ----- -----
+  Model (error reduced) | 323.364 14 23.097  618.850 .9914 .0000
+      A                 | 141.497  2 70.749 1895.568 .9806 .0000
+      B                 | 181.500  4 45.375 1215.734 .9848 .0000
+    A:B                 |   0.367  8  0.046    1.229 .1159 .2943
+  Error (from model)    |   2.799 75  0.037
+  ----- --------------- | ------- -- ------ -------- ----- -----
+  Total (empty model)   | 326.163 89  3.665
```

How does R Handle Such Models ?

```
> supernova(lm(X~A+A:B, data=FDdat))

+  Analysis of Variance Table (Type III SS)
+  Model: X ~ A + A:B
+
+                             SS df     MS       F   PRE     p
+  ----- --------------- | ------- -- ------ ------- ----- -----
+  Model (error reduced) | 323.364 14 23.097 618.850 .9914 .0000
+      A                 |  27.853  2 13.927 373.135 .9087 .0000
+    A:B                 | 181.867 12 15.156 406.064 .9848 .0000
+  Error (from model)    |   2.799 75  0.037
+  ----- --------------- | ------- -- ------ ------- ----- -----
+  Total (empty model)   | 326.163 89  3.665

> supernova(lm(X~A+A:B, data=FDdat), type=2)

+  Analysis of Variance Table (Type II SS)
+  Model: X ~ A + A:B
+
+                             SS df     MS        F   PRE     p
+  ----- --------------- | ------- -- ------ -------- ----- -----
+  Model (error reduced) | 323.364 14 23.097  618.850 .9914 .0000
+      A                 | 141.497  2 70.749 1895.568 .9806 .0000
+    A:B                 | 181.867  4 45.467 1218.192 .9848 .0000
+  Error (from model)    |   2.799 75  0.037
+  ----- --------------- | ------- -- ------ -------- ----- -----
+  Total (empty model)   | 326.163 89  3.665
```

How does R Handle Such Models ?

Anova from the car package

```
> library(car)
> Anova(lm(X~A+B+A:B, data=FDdat))

+ Anova Table (Type II tests)
+
+ Response: X
+            Sum Sq Df   F value Pr(>F)
+ A         141.497  2 1895.5681 <2e-16 ***
+ B         181.500  4 1215.7341 <2e-16 ***
+ A:B         0.367  8    1.2288 0.2943
+ Residuals   2.799 75
+ ---
+ Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

How does R Handle Such Models ?

```
> Anova(lm(X~A+B+A:B, data=FDdat), type="III")

+ Anova Table (Type III tests)
+
+ Response: X
+             Sum Sq Df  F value Pr(>F)
+ (Intercept)  7.402  1 198.3253 <2e-16 ***
+ A           27.853  2 373.1349 <2e-16 ***
+ B           57.798  4 387.1432 <2e-16 ***
+ A:B           0.367  8   1.2288 0.2943
+ Residuals     2.799 75
+ ---
+ Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

```
> Anova(lm(X~-1+A+B+A:B, data=FDdat), type="III")

+ Anova Table (Type III tests)
+
+ Response: X
+            Sum Sq Df   F value Pr(>F)
+ A         161.454  3 1441.9444 <2e-16 ***
+ B          57.798  4  387.1432 <2e-16 ***
+ A:B         0.367  8    1.2288 0.2943
+ Residuals   2.799 75
+ ---
+ Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

How does R Handle Such Models ?

This analysis tells us that

- `aov` and `Anova` report the Type II sums of squares by default in these multifactor settings
- `supernova` reports the Type III sums of squares

How does R Handle Such Models ?

Note: for models without interactions, Type II and Type III sums of squares are `identical` for the balanced design.

```
> supernova(lm(X~A+B, data=FDdat), type=2)

+  Analysis of Variance Table (Type II SS)
+  Model: X ~ A + B
+
+                              SS df     MS        F   PRE     p
+  ----- --------------- | ------- -- ------ -------- ----- -----
+  Model (error reduced) | 322.997  6 53.833 1411.221 .9903 .0000
+      A                 | 141.497  2 70.749 1854.664 .9781 .0000
+      B                 | 181.500  4 45.375 1189.500 .9829 .0000
+  Error (from model)    |   3.166 83  0.038
+  ----- --------------- | ------- -- ------ -------- ----- -----
+  Total (empty model)   | 326.163 89  3.665
```

How does R Handle Such Models ?

```
> supernova(lm(X~A+B, data=FDdat), type=3)

+  Analysis of Variance Table (Type III SS)
+  Model: X ~ A + B
+
+                              SS df     MS        F   PRE     p
+  ----- --------------- | ------- -- ------ -------- ----- -----
+  Model (error reduced) | 322.997  6 53.833 1411.221 .9903 .0000
+      A                 | 141.497  2 70.749 1854.664 .9781 .0000
+      B                 | 181.500  4 45.375 1189.500 .9829 .0000
+  Error (from model)    |   3.166 83  0.038
+  ----- --------------- | ------- -- ------ -------- ----- -----
+  Total (empty model)   | 326.163 89  3.665
```

Example: Batteries Data

A - Material

B - Temperature

Model A + B + A:B

```
> summary(aov(life~material*temp,data=Batteries))

+              Df Sum Sq Mean Sq F value      Pr(>F)
+ material      2  10633    5317   7.983     0.00189 **
+ temp          2  39083   19542  29.344 0.000000169 ***
+ material:temp  4   9438    2359   3.543     0.01897 *
+ Residuals     27  17981     666
+ ---
+ Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

$$SS = SST_A + SST_B + SSI_{AB} + SSE$$

Example: Batteries Data

A - Material

B - Temperature

Model A + A:B

```
> summary(aov(life~material+material:temp,data=Batteries))

+               Df Sum Sq Mean Sq F value     Pr(>F)
+ material       2  10633    5317   7.983    0.00189 **
+ material:temp  6  48521    8087  12.143 0.00000135 ***
+ Residuals     27  17981     666
+ ---
+ Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

$$SS = SST_A + SSI_{B:AB} + SSE$$

where

$$SSI_{B:AB} = SST_B + SSI_{AB}$$

Example: Batteries Data

A - Material

B - Temperature

Model B + A:B

```
> summary(aov(life~temp+material:temp,data=Batteries))

+               Df Sum Sq Mean Sq F value      Pr(>F)
+ temp           2  39083   19542  29.344 0.000000169 ***
+ temp:material  6  20071    3345   5.023     0.00142 **
+ Residuals     27  17981     666
+ ---
+ Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

$$SS = SST_B + SSI_{A:AB} + SSE$$

where

$$SSI_{A:AB} = SST_A + SSI_{AB}$$

Example: Batteries Data

A - Material

B - Temperature

Model A:B

```
> summary(aov(life~material:temp,data=Batteries))

+               Df Sum Sq Mean Sq F value      Pr(>F)
+ material:temp  8  59154    7394    11.1 0.000000861 ***
+ Residuals     27  17981     666
+ ---
+ Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

$$SS = SSI_{A:B:AB} + SSE$$

where

$$SSI_{A:B:AB} = SST_A + SST_B + SSI_{AB}$$

Final Note on ANOVA

We have studied the simplest design scenarios: extension to

- incomplete
- unbalanced
- nested
- random effect

designs are possible, but more complicated statistically.

Furthermore R has greater functionality: for example, it has the capability to carry out ANOVA-like analyses even for the case of non-equal variances (when Levene's test rejects the hypothesis of equal variances).

Chapter 2
Linear Regression Modelling

In this section

- Textbook Chapters 11 & 12
- Straight-line regression
- Polynomial regression
- Multiple regression
- Factor predictors
- The Linear Model
- Model selection and checking

Linear regression Modelling

In the previous section, we attempted to explain the variation in an observed response variable by fitting models with one or more factors.

Factors are discrete variables taking different levels; in this section we will now utilize continuous variables that can similarly explain variation in an observed response.

Simple Linear Regression

We will investigate models relating two quantities x and y through equations of the form

$$y = ax + b$$

where a and b are constants (that is, a straight line).

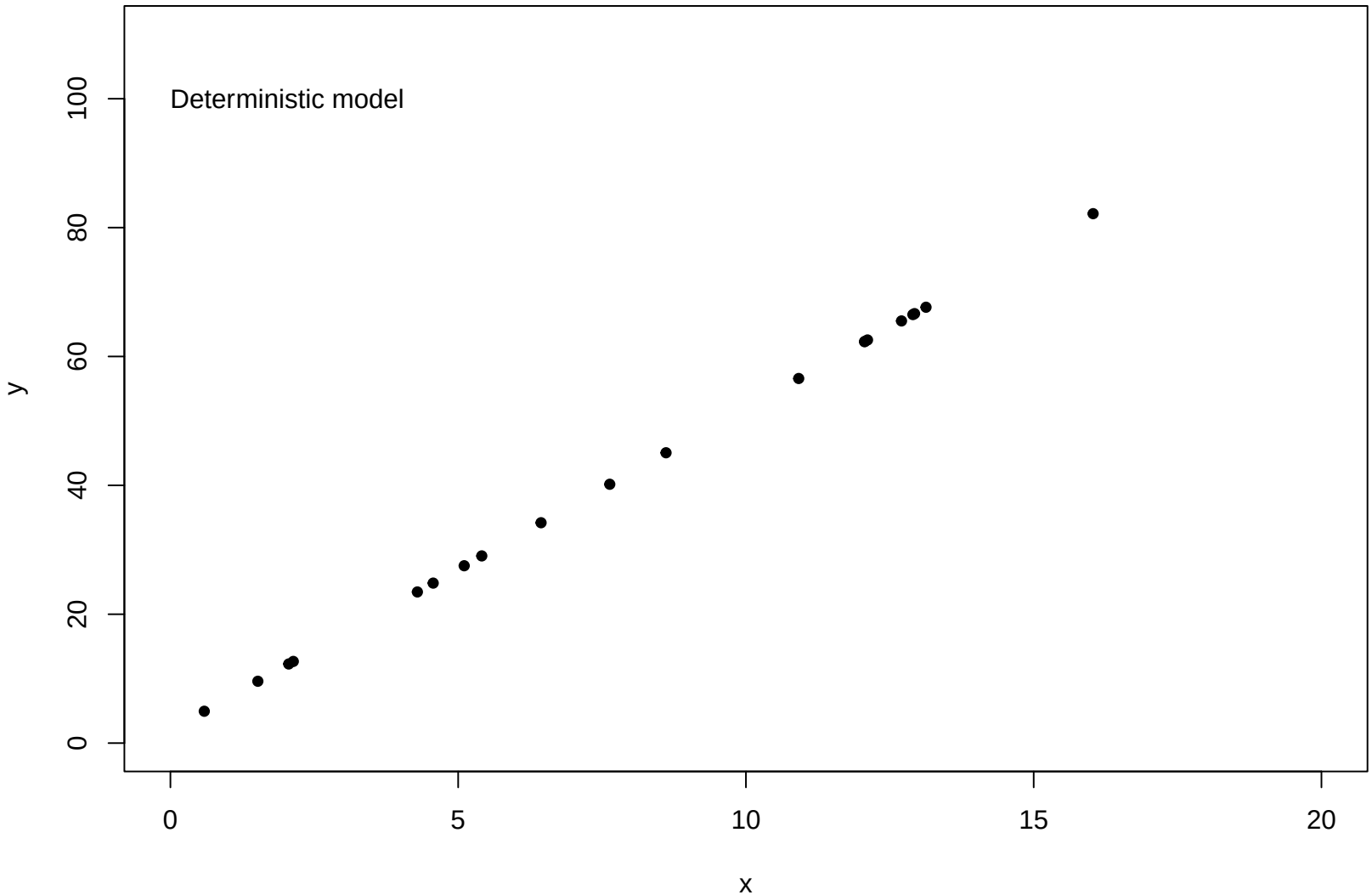
Variables x and y will not be treated exchangeably - we will regard y as being a function of x.

Simple Linear Regression

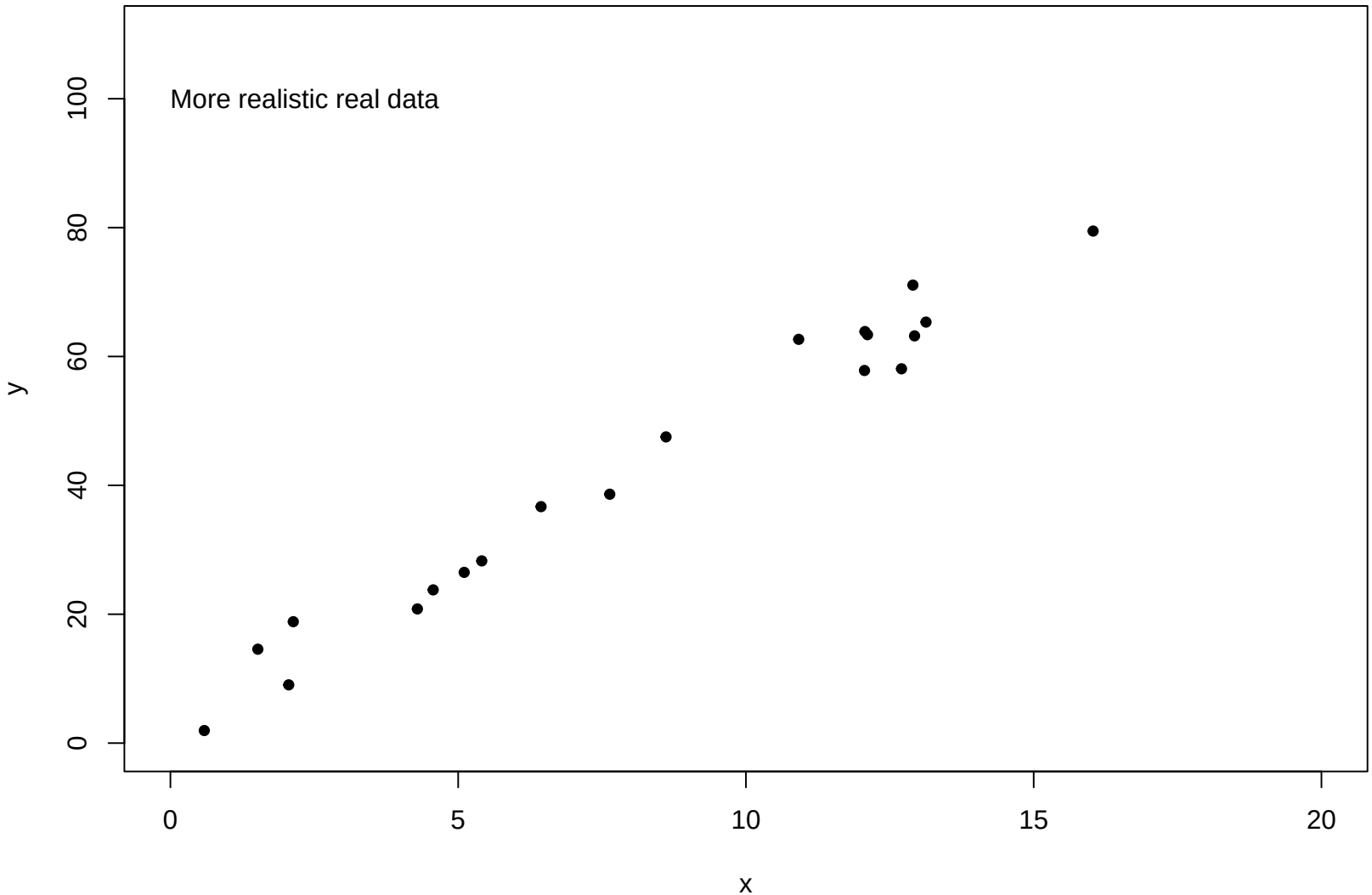
Such models are deterministic, that is, if we know x (and the values of the constants), we can compute y exactly without error.

A more useful model allows for the possibility that the system is not observed perfectly, that is, we do not observe (x, y) pairs that are always consistent with a simple functional relationship.

Simple Linear Regression



Simple Linear Regression



Simple Linear Regression

Example: Pharmacokinetic Model

If a dose of drug is taken at time $x = 0$, the concentration of drug still in the bloodstream at time x is often well-modelled by a simple equation.

Let

- D denote the amount of drug taken at $x = 0$
- x time
- $y^\star$ is the concentration in the bloodstream.

Simple Linear Regression

Example: Pharmacokinetic Model

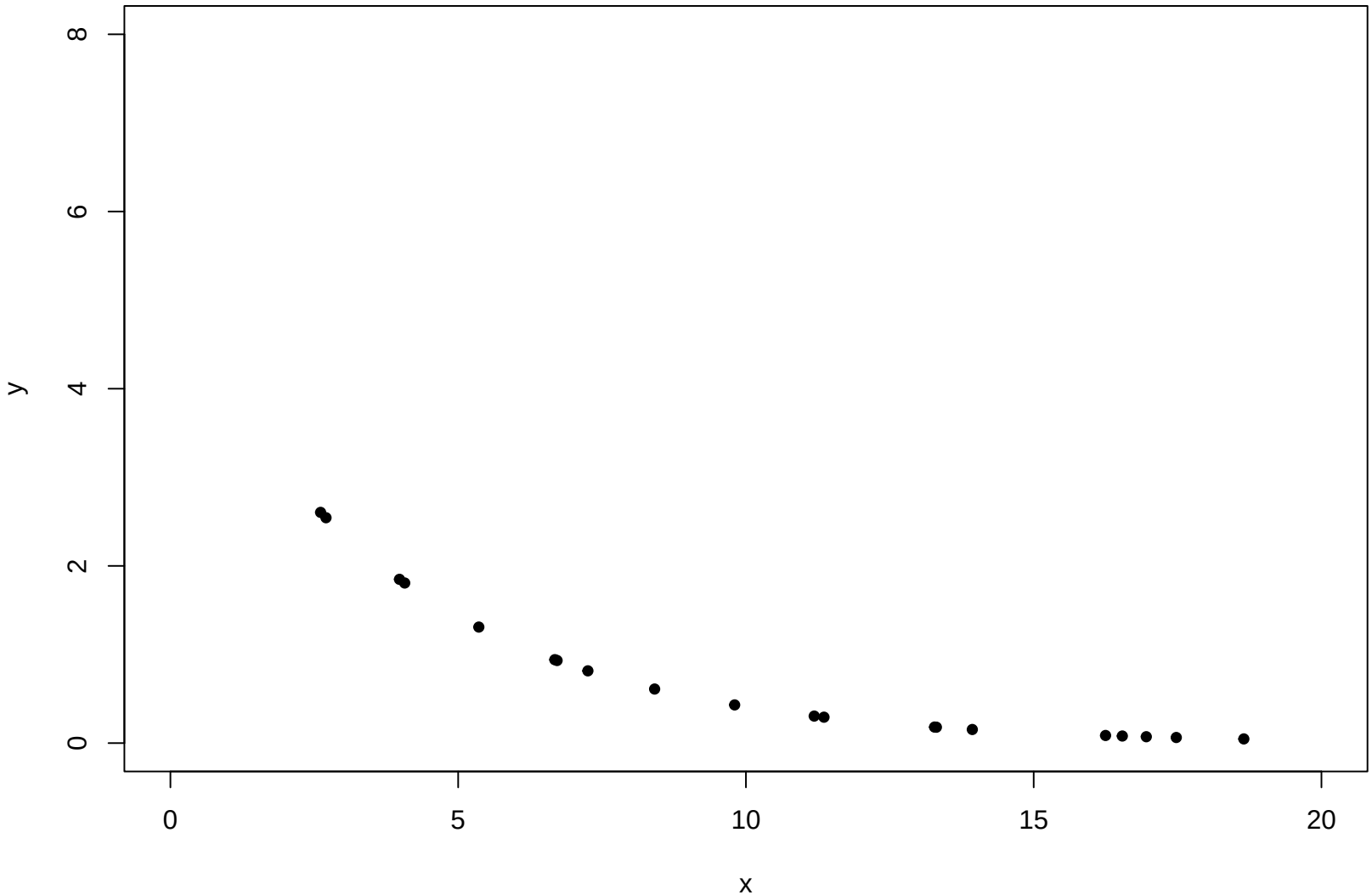
Then

$$y^{\star} = \frac{D}{V} \exp\{-\lambda x\}$$

where

- λ is the elimination rate
- V is the volume of bloodstream.

Simple Linear Regression



Simple Linear Regression

Example: Pharmacokinetic Model

Taking logs of both sides, setting $y = \log y^\star$, then

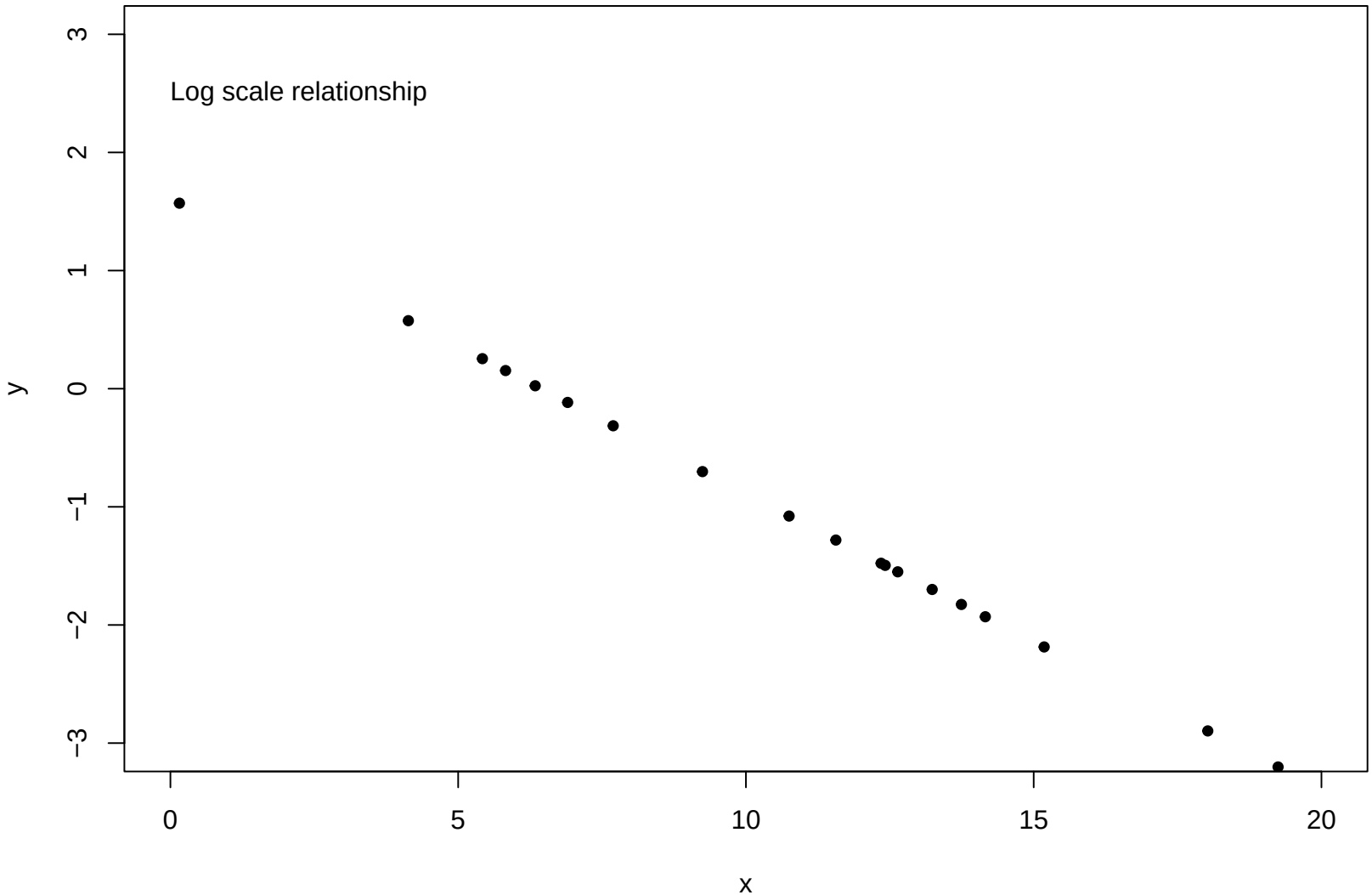
$$y = -\lambda x + \log(D/V) = -\lambda x + (\log D - \log V)$$

that is, $y = ax + b$ where

- $a = -\lambda$
- $b = (\log D - \log V)$

Here $\log$ means natural log $\ln$.

Simple Linear Regression

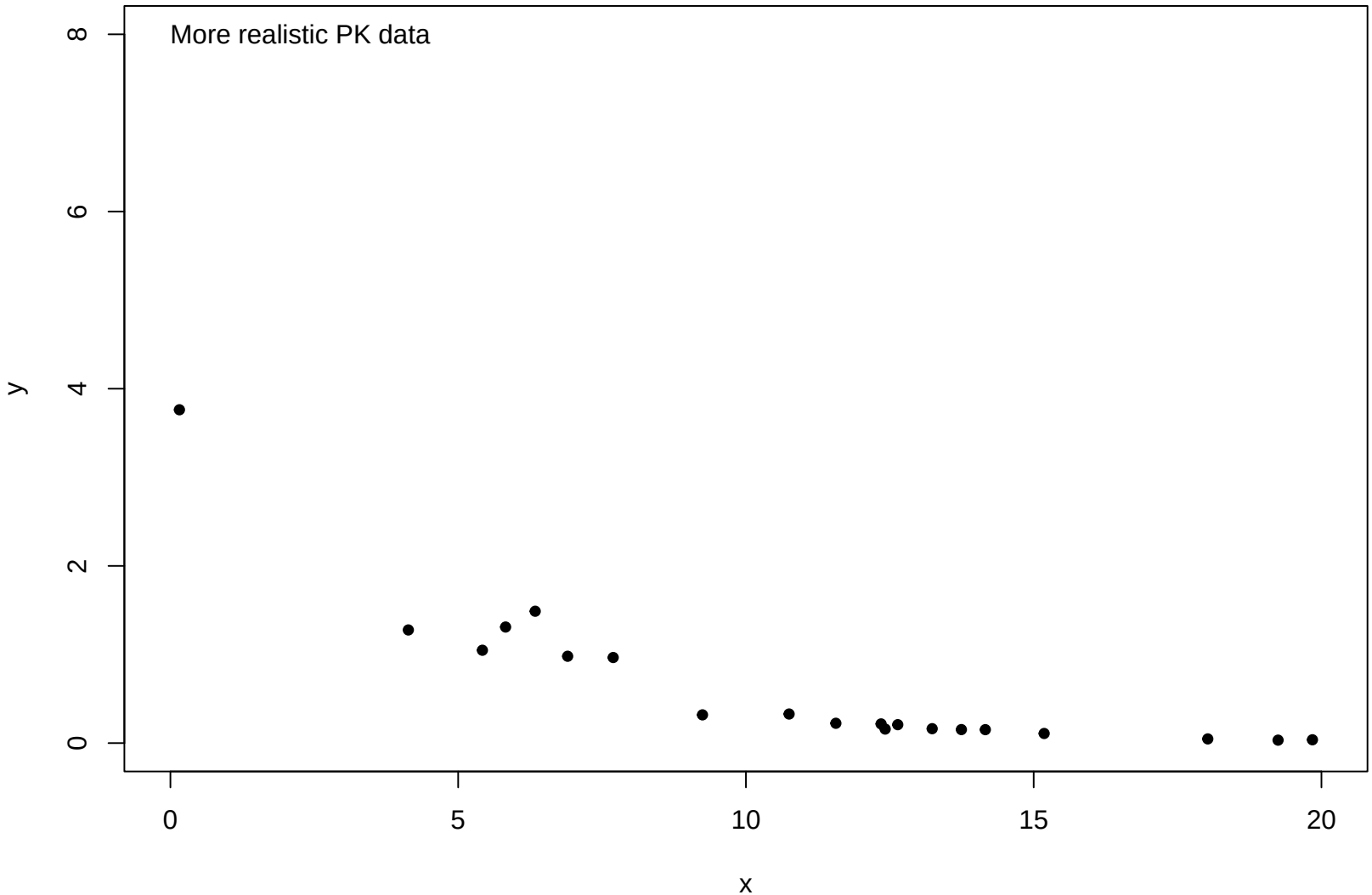


Simple Linear Regression

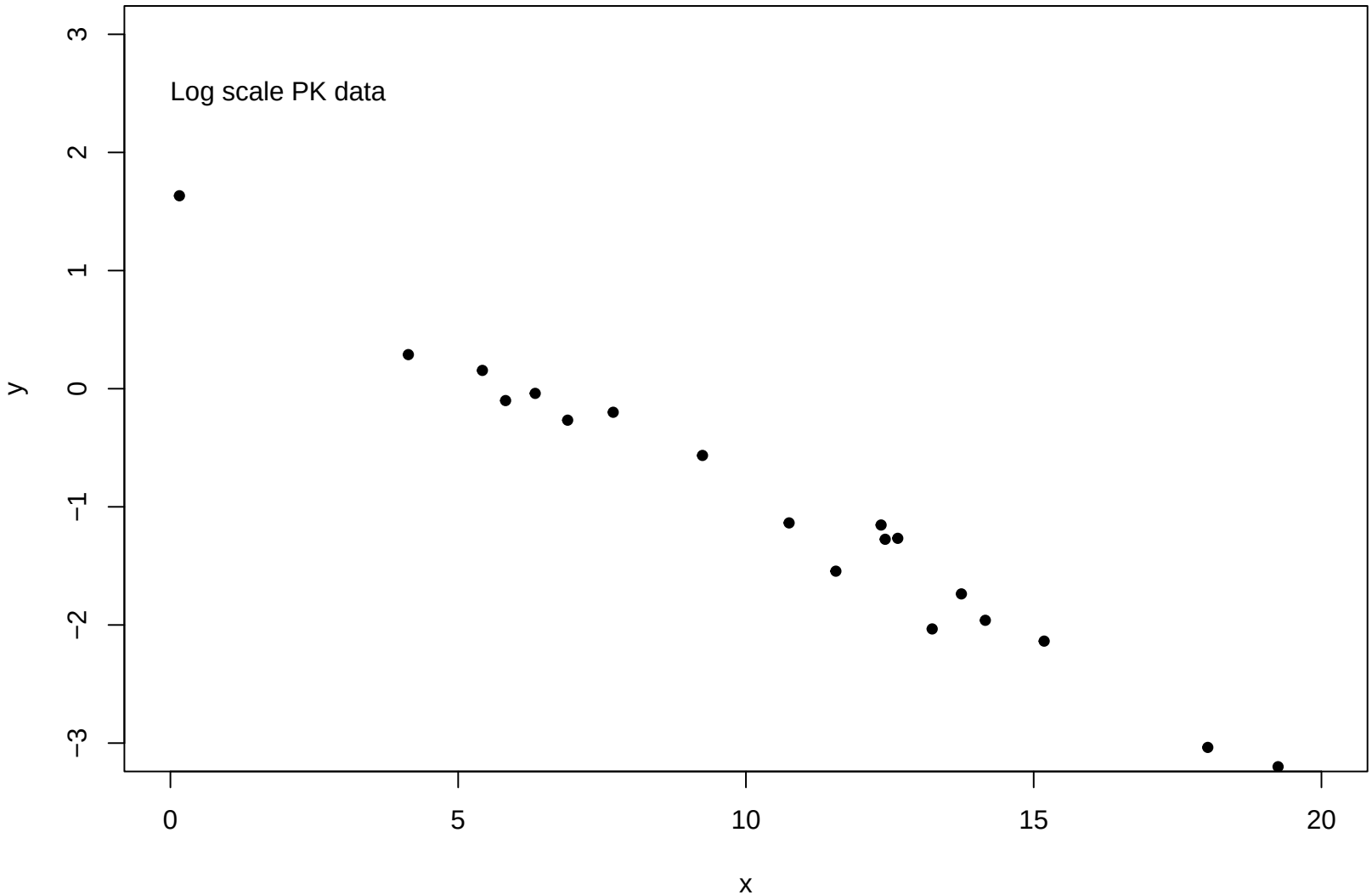
However, in practice, when we measure concentration, we do so with random error.

- We need to acknowledge the presence of random error in the measurement process
- Deterministic models do not permit this

Simple Linear Regression



Simple Linear Regression



Probabilistic Models

In a probabilistic model, we allow for the possibility that y is observed with random error, that is,

$$y = ax + b + ERROR$$

where $ERROR$ is a random term that is present due to imperfect observation of the system due to measurement error or incomplete information.

Probabilistic Models

Note that we do not treat x and y interchangeably;

- x is a fixed quantity that is measured without error
- y is measured with random error.

A Basic Probabilistic Model

We model the variation in y as a function of x. We observe data pairs $(x_i, y_i), i = 1, \ldots, n$.

Terminology:

- y - Dependent variable or response variable
- x - Independent variable, or predictor, or covariate

A Basic Probabilistic Model

The model we study takes the form

$$y = \beta_0 + \beta_1 \mathbf{x} + \epsilon$$

where ϵ is a random error term, specifically

- a random variable with mean zero and finite variance σ^2
- represents the error present in the measurement of the response.

Parameters:

- β_0 - Intercept parameter
- β_1 - Slope parameter

A Basic Probabilistic Model

For example

- $\beta_1 > 0$ - increasing y with increasing x
- $\beta_1 < 0$ - decreasing y with increasing x
- $\beta_1 = 0$ - no relationship between x and y

Note:

$$\mathbb{E}[Y|x] = \beta_0 + \beta_1 x$$

where $\mathbb{E}[Y|x]$ is the expected value of Y for a fixed value of x.

A Basic Probabilistic Model

Recall the notation

- Y - a random variable with a probability distribution
- y - a fixed observed value of the variable Y.

A Basic Probabilistic Model

Fundamental Problem: If we believe the straight-line model with error is correct, how do we find the values of parameters β_0 and β_1.

We only have the observed data $\{(x_i, y_i), i = 1, \ldots, n\}$.

Least Squares Fitting

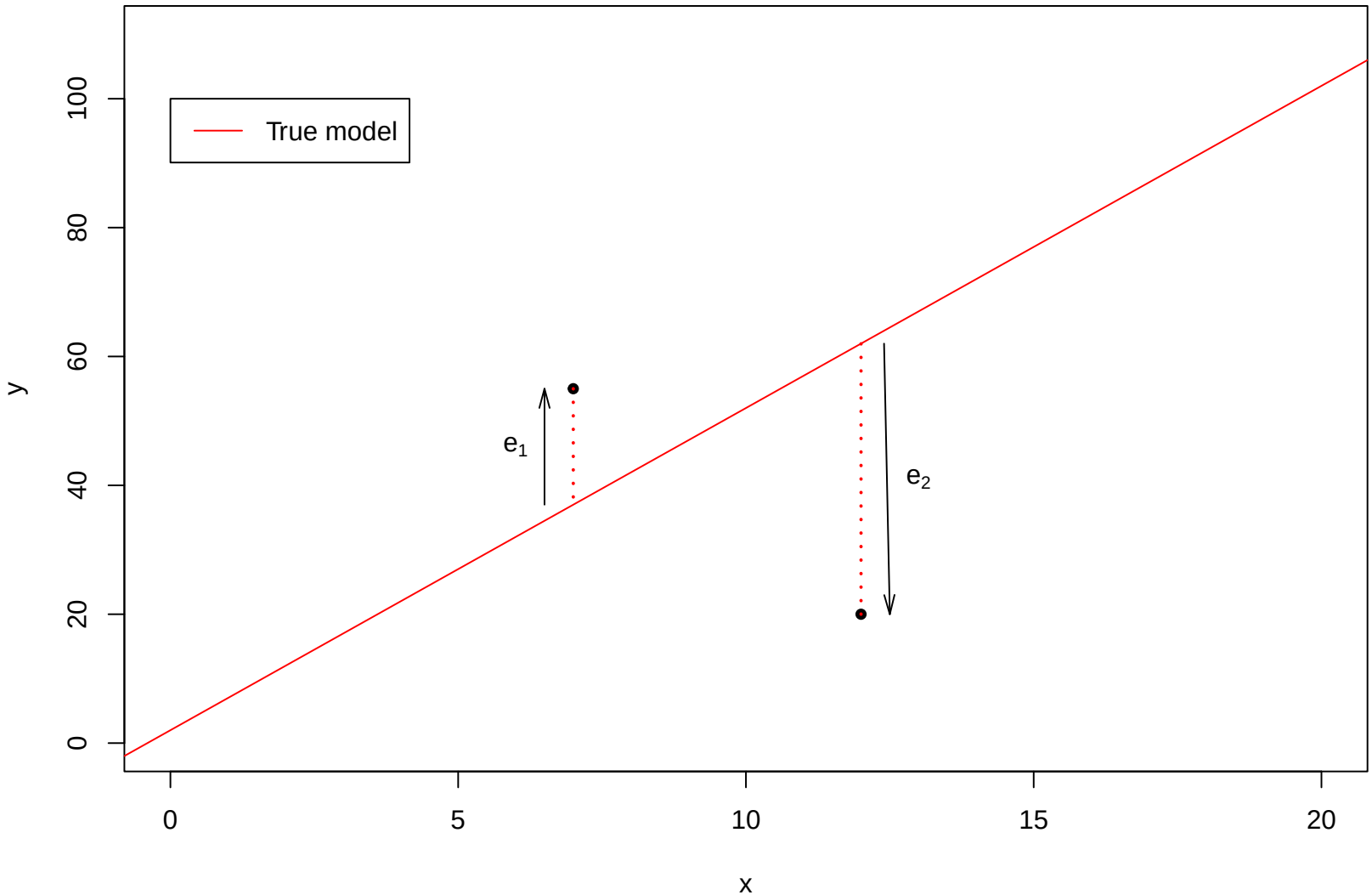
We select the best values of β_0 and β_1 by minimizing the error in fit. For two data pairs (x_1, y_1) and (x_2, y_2), the errors in fit are

$$\begin{aligned} e_1 &= y_1 - (\beta_0 + \beta_1 x_1) \\ e_2 &= y_2 - (\beta_0 + \beta_1 x_2) \end{aligned}$$

respectively.

With two points, we can interpolate the data with a straight line.

Least Squares Fitting



Least Squares Fitting

It is reasonable to choose the best parameters by considering the aggregate errors across the data set: we might inspect

$$\sum_{i=1}^{2} e_i = e_1 + e_2$$

and seek to minimize this by manipulating (β_0, β_1).

But note that, potentially, $e_1 > 0$ and $e_2 < 0$ so there is a possibility that these fitting errors cancel each other out.

Least Squares Fitting

Therefore we look at squared errors (as a large negative error is as bad as a large positive error)

$$\begin{aligned} e_1^2 &= (y_1 - (\beta_0 + \beta_1 x_1))^2 \\ e_2^2 &= (y_2 - (\beta_0 + \beta_1 x_2))^2 \end{aligned}$$

Least Squares Fitting

For n data, we obtain n misfit squared errors

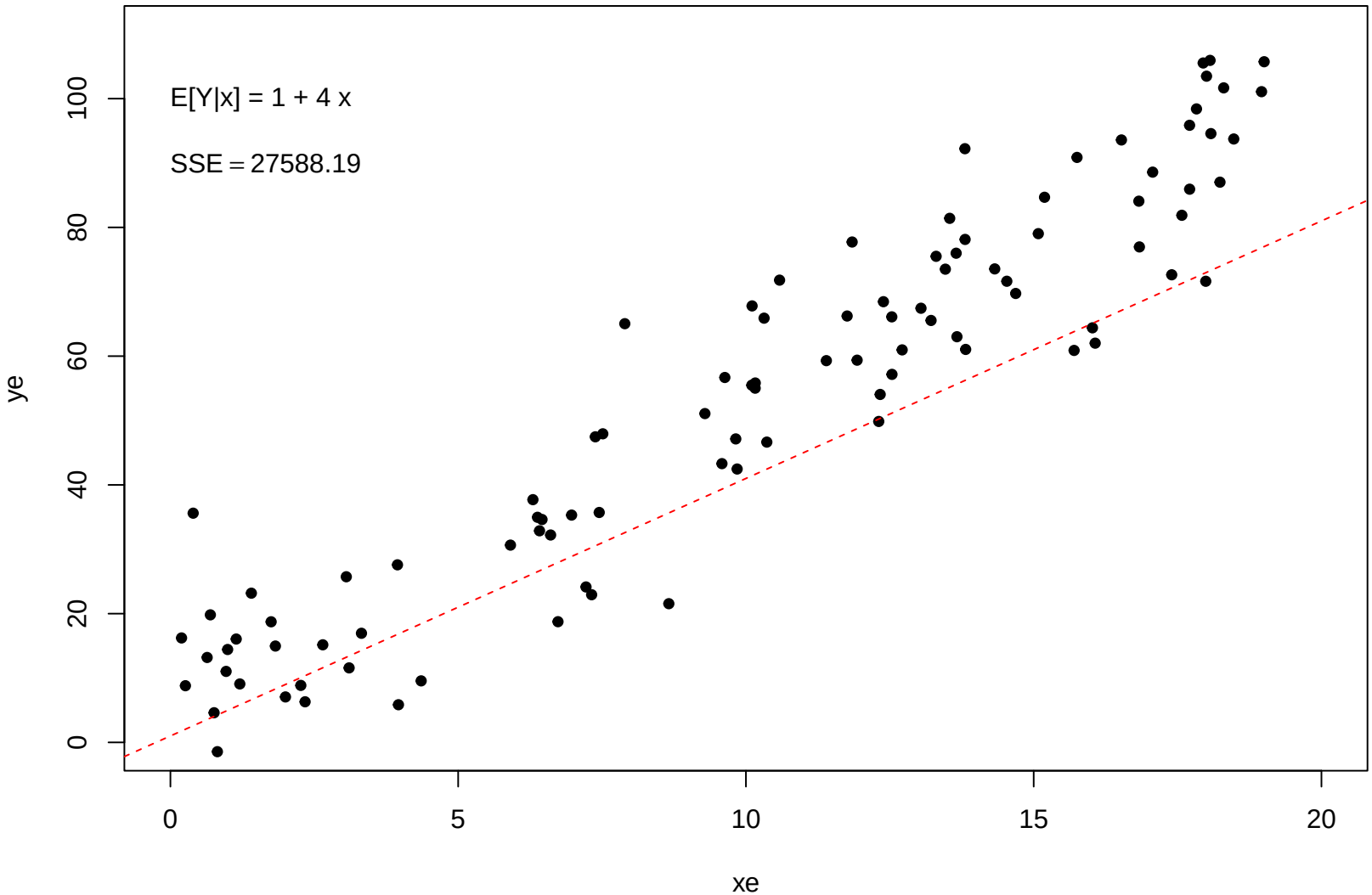
$$e_1^2, \ldots, e_n^2$$

We select β_0 and β_1 as the values of the parameters that minimize SSE, where

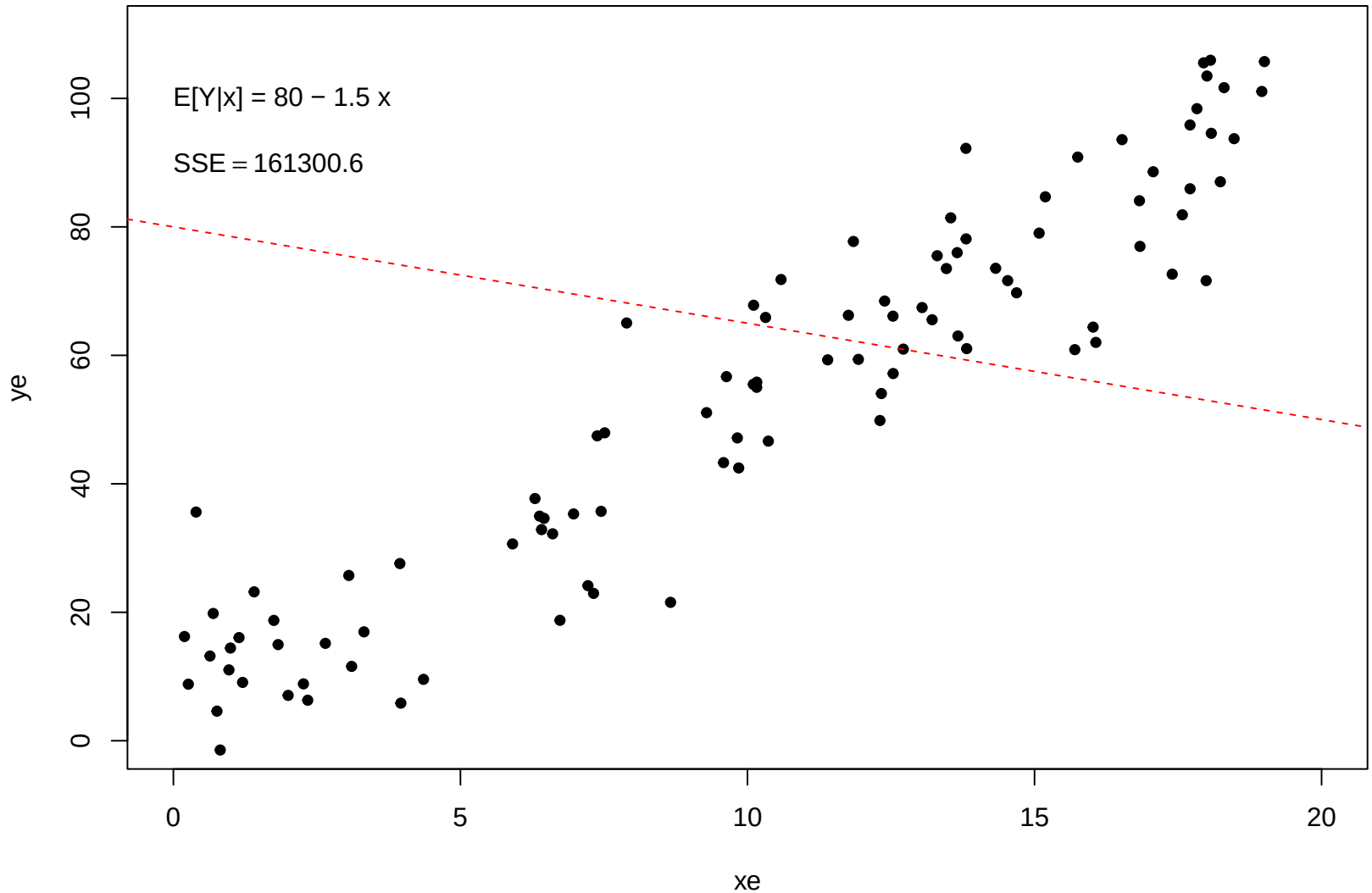
$$SSE = \sum_{i=1}^{n} e_i^2 = \sum_{i=1}^{n} (y_i - (\beta_0 + \beta_1 x_i))^2$$

We wish to make the total misfit squared error as small as possible.

Least Squares Fitting



Least Squares Fitting



Least Squares Fitting

SSE - sum of squared errors - is similar to the SSE for ANOVA. We could write

$$SSE = SSE(\beta_0, \beta_1)$$

to show the dependence of SSE on the parameters.

Minimization of $SSE(\beta_0, \beta_1)$ is achieved analytically using calculus.

Least Squares Fitting

It follows that the best-fit values $\widehat{\beta}_0$ and $\widehat{\beta}_1$ are given by

$$\widehat{\beta}_1 = \frac{SS_{xy}}{SS_{xx}} \qquad \widehat{\beta}_0 = \overline{y} - \widehat{\beta}_1 \overline{x}$$

where

- SS_{xx}:

$$SS_{xx} = \sum_{i=1}^{n} (x_i - \overline{x})^2$$

- SS_{xy}:

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

Least Squares Fitting

$\widehat{\beta}_0$ and $\widehat{\beta}_1$ are the least-squares estimates

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

is the least-squares line of best fit. The fitted-values are

$$\widehat{y}_i = \widehat{\beta}_0 + \widehat{\beta}_1 x_i \qquad i = 1, \ldots, n$$

and the residuals are

$$\widehat{e}_i = y_i - \widehat{y}_i = y_i - \widehat{\beta}_0 - \widehat{\beta}_1 x_i \qquad i = 1, \ldots, n$$

Estimating σ^2

Using the LS procedure, we can construct an estimate of the error or residual error variance

Recall that the variance of ϵ is σ^2, and an estimate of σ^2 is

$$\widehat{\sigma}^2 = \frac{SSE(\widehat{\beta}_0, \widehat{\beta}_1)}{n-2} = s^2$$

say.

Estimating σ^2

Note that the denominator $n-2$ is again a degrees of freedom parameter of the form

(Number of data points) - (Number of model parameters)

or $n - p$. Here $p = 2$ (β_0 and β_1).

Note also that

$$SSE(\widehat{\beta}_0, \widehat{\beta}_1) = \sum_{i=1}^{n} (y_i - \widehat{y}_i)^2 = SS_{yy} - \widehat{\beta}_1 SS_{xy}$$

where

$$SS_{yy} = \sum_{i=1}^{n} (y_i - \overline{y})^2$$

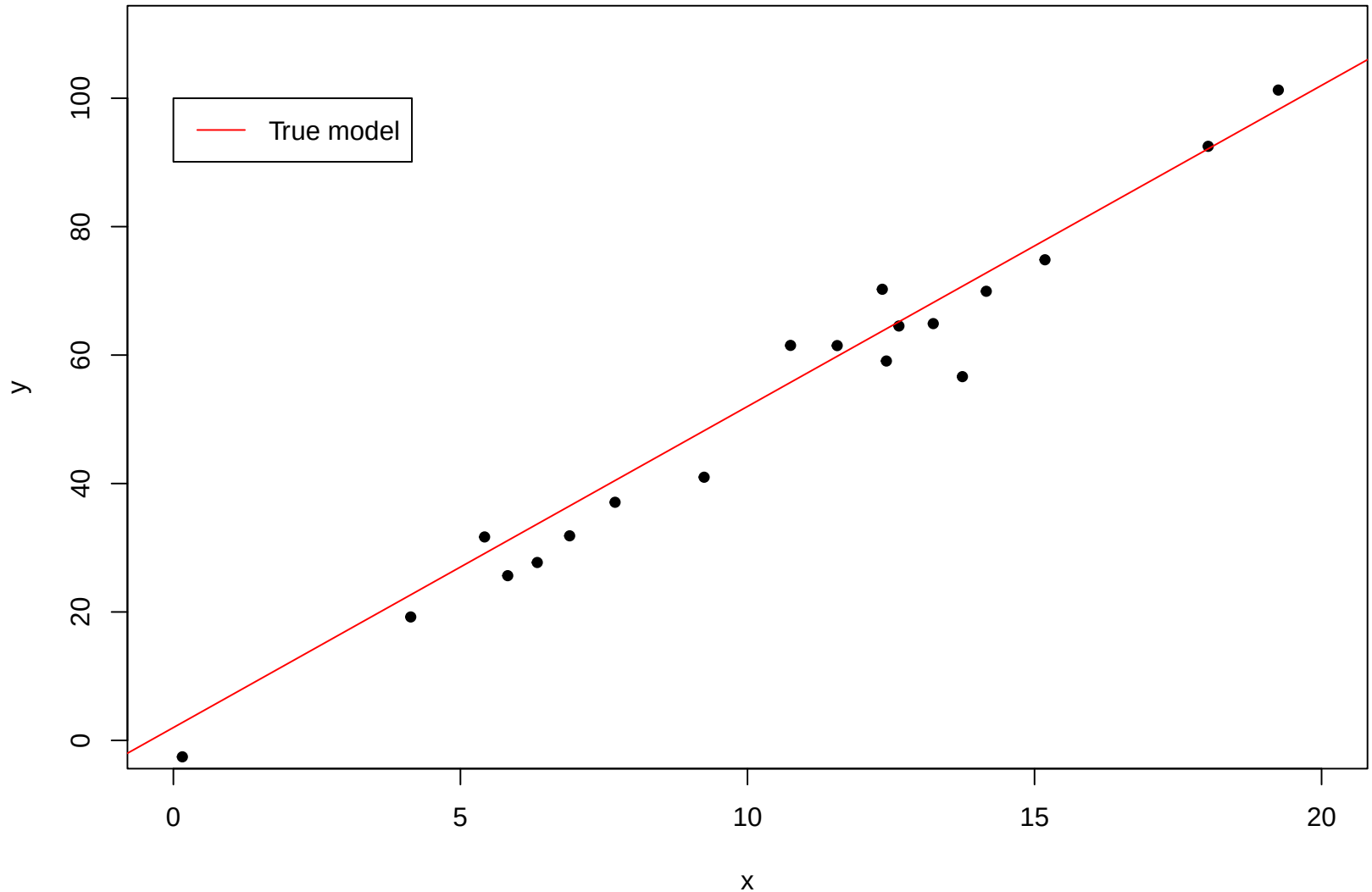
Example

Simulated example: here $n = 20$ and

$$\beta_0 = 2 \qquad \beta_1 = 5 \qquad \sigma^2 = 6^2$$

```
> set.seed(1234)
> n<-20
> y<-2+5*x+rnorm(n)*6
> par(mar=c(4,4,2,2))
> plot(x,y,pch=19,cex=0.8,xlim=range(0,20), ylim=range(0,110))
> abline(2,5,col='red')
> legend(0,100,c('True model'),col='red',lty=1)
```

Example



Example

Estimation:

```
> SS.xx<-sum((x-mean(x))^2)
> SS.xy<-sum((x-mean(x))*(y-mean(y)))
> b1.hat<-SS.xy/SS.xx
> b0.hat<-mean(y)-b1.hat*mean(x)
> sigsq.hat<-sum((y - b0.hat - b1.hat*x)^2)/(n-2)
> c(b0.hat,b1.hat,sigsq.hat)

+ [1] -4.851720  5.488668 32.190393
```

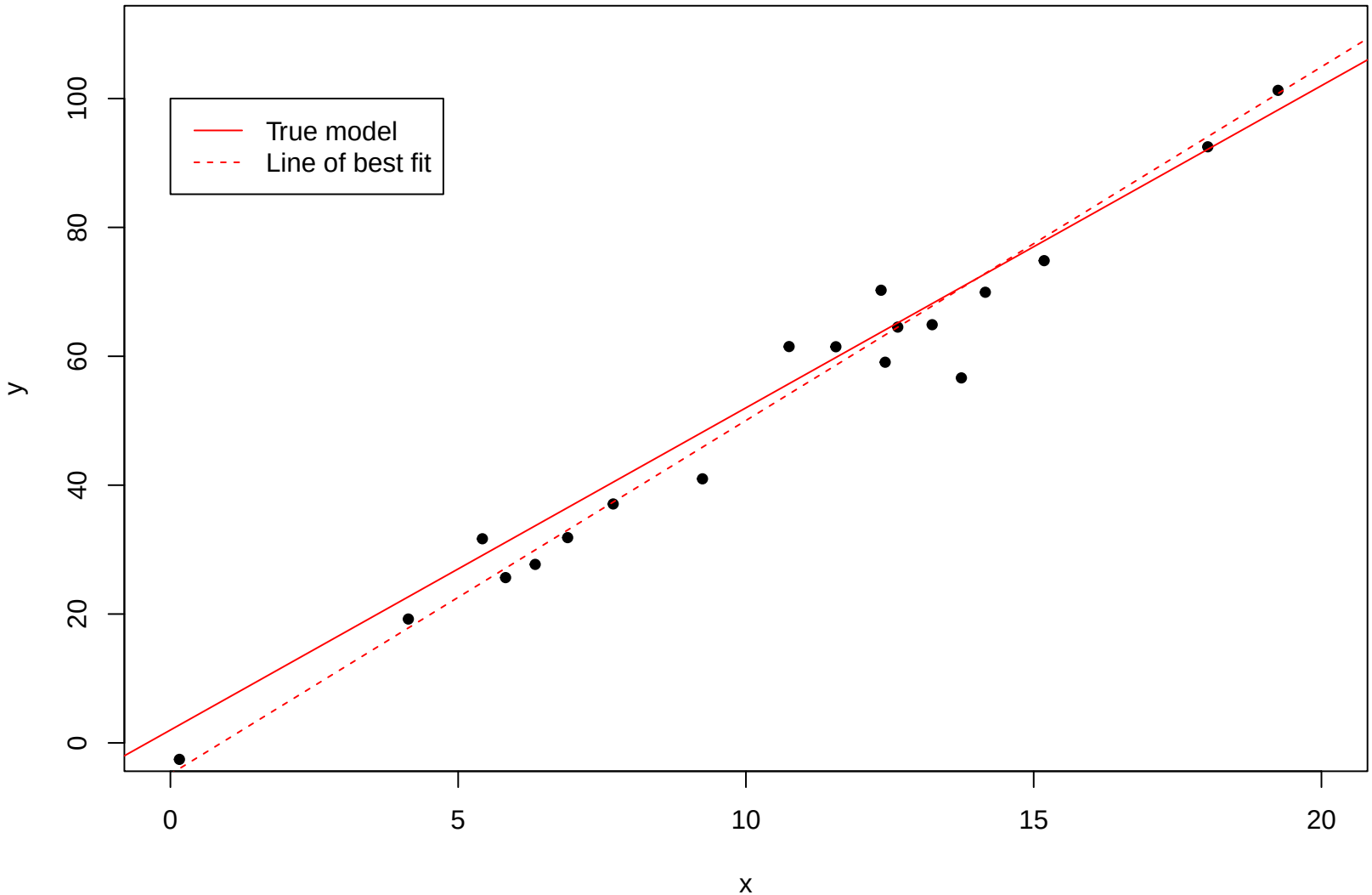
That is

$$\widehat{\beta}_0 = -4.8517 \qquad \widehat{\beta}_1 = 5.4887 \qquad \widehat{\sigma}^2 = 32.1904$$

so that

$$\widehat{\sigma} = 5.6737.$$

Example



Example

Using the `lm` function

```
> fit1<-lm(y~x)
> summary(fit1)


+
+ Call:
+ lm(formula = y ~ x)
+
+ Residuals:
+      Min       1Q   Median       3Q      Max
+ -13.9386  -2.9048  -0.7366   1.7927  11.6499
+
+ Coefficients:
+             Estimate Std. Error t value Pr(>|t|)
+ (Intercept)  -4.8517     3.0097  -1.612    0.124
+ x             5.4887     0.2494  22.008 1.84e-14 ***
+ ---
+ Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
+
+ Residual standard error: 5.674 on 18 degrees of freedom
+ Multiple R-squared:  0.9642, Adjusted R-squared:  0.9622
+ F-statistic: 484.4 on 1 and 18 DF,  p-value: 1.835e-14
```

Model Assumptions for Least-Squares

To utilize least-squares (LS) for the probabilistic model

$$Y = \beta_0 + \beta_1 \mathbf{x} + \epsilon$$

we make the following four assumptions

1. The expected value of ϵ is zero so that

$$\mathbb{E}[Y|\mathbf{x}] = \beta_0 + \beta_1 \mathbf{x}.$$

2. The variance of ϵ, denoted σ^2, is constant and does not depend on x.

Model Assumptions for Least-Squares

3. The probability distribution of ϵ is a symmetric distribution about zero
    - a stronger assumption is that $\epsilon \sim Normal(0, \sigma^2)$.
4. The errors for two different measured responses at x_1 and x_2 are independent
    - the error ϵ_1 in measuring y_1 at x_1 is independent of the error ϵ_2 in measuring y_2 at x_2.

Hypothesis Testing

In the model where

$$\mathbb{E}[Y|\mathbf{x}] = \beta_0 + \beta_1 \mathbf{x}$$

it is of interest to test the hypothesis

$$\begin{aligned} H_0 &: \beta_1 = 0 \\ H_a &: \beta_1 \neq 0 \end{aligned}$$

H_0 implies that there is no systematic influence of x on the value of Y. This is a two-sided alternative.

Hypothesis Testing

To test H_0 vs H_a we us the test statistic

$$T = \frac{\widehat{\beta}_1}{\text{e.s.e}(\widehat{\beta}_1)}$$

where $\text{e.s.e}(\widehat{\beta}_1)$ is the Estimated Standard Error of $\widehat{\beta}_1$, computed as

$$\text{e.s.e}(\widehat{\beta}_1) = \frac{s}{\sqrt{SS_{xx}}}$$

where s is the estimate of σ defined previously.

Hypothesis Testing

If H_0 is true, and in fact $\beta_1 = 0$, then

$$T = \frac{\widehat{\beta}_1}{s/\sqrt{SS_{xx}}} \sim t_{n-2}$$

that is, the t-distribution with $n - 2$ degrees of freedom.

We can carry out a significance test at level α in the usual way (use a p-value, or construct the rejection region).

Hypothesis Testing

We might also consider a one-sided test:

- If $H_a : \beta_1 \neq 0$, we use the two-sided rejection region, with critical values
$$C_R = \pm t_{\alpha/2, n-2}$$
- If $H_a : \beta_1 > 0$, we use the one-sided rejection region, with critical value
$$C_R = t_{\alpha, n-2}$$
- If $H_a : \beta_1 < 0$, we use the one-sided rejection region, with critical value
$$C_R = -t_{\alpha, n-2}$$

Hypothesis Testing

To test

$$\begin{aligned} H_0 &: \beta_1 = b \\ H_a &: \beta_1 \neq b \end{aligned}$$

for any b, the test statistic is

$$T = \frac{\widehat{\beta}_1 - b}{s/\sqrt{SS_{xx}}}$$

(for example, $b = 1$ may be of interest). If H_0 is true

$$T \sim t_{n-2}.$$

Hypothesis Testing

`lm` computes the key quantities. Using the `summary` function

```
> summary(fit1)$coef

+              Estimate Std. Error   t value     Pr(>|t|)
+ (Intercept) -4.851720  3.0096733 -1.612042 1.243479e-01
+ x            5.488668  0.2493914 22.008246 1.835225e-14
```

- first column: estimates
- second column: estimated standard errors
- third column: test statistic where null hypothesis is that parameter is zero
- fourth column: two-sided p-value for the test

Confidence Interval

A $100(1-\alpha)\%$ confidence interval for β_1 is

$$\widehat{\beta}_1 \pm t_{\alpha/2,n-2} \times \text{e.s.e.}(\widehat{\beta}_1)$$

where

$$\begin{aligned} t_{\alpha/2,n-2} &: \text{upper } \alpha/2 \text{ point of } t_{n-2} \text{ distn} \\ \text{e.s.e.}(\widehat{\beta}_1) &: \text{estimated standard error of } \widehat{\beta}_1 \end{aligned}$$

Confidence Interval

Note: we could perform a similar analysis for β_0, but this is generally of less interest.

The only quantity that needs attention is the estimated standard error of $\widehat{\beta}_0$. It can be shown that

$$\text{e.s.e.}(\widehat{\beta}_0) = \sqrt{\frac{1}{n}\left(1 + \frac{n\overline{x}^2}{SS_{xx}}\right)}$$

Confidence Interval

The `confint` function computes the confidence intervals:

```
> confint(fit1)

+                 2.5 %   97.5 %
+ (Intercept) -11.174810 1.471369
+ x             4.964716 6.012620
```

That is, the 95% CIs for β_0 and β_1 are

$$(-11.174810, 1.471369) \qquad \text{and} \qquad (4.964716, 6.012620)$$

respectively.

The Coefficient of Correlation

To measure the strength of association between the two variables x and y we can use the

Pearson Product Moment Coefficient Of Correlation

or correlation coefficient which measures the strength of the linear relationship between x and y.

The Coefficient of Correlation

The correlation coefficient, r, is defined by

$$r = \frac{SS_{xy}}{\sqrt{SS_{xx}SS_{yy}}}$$

where

$$SS_{xx} = \sum_{i=1}^{n} (x_i - \overline{x})^2 \qquad SS_{yy} = \sum_{i=1}^{n} (y_i - \overline{y})^2$$

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

The Coefficient of Correlation

In R the `cor` function can be used

```
> cor(x,y)

+ [1] 0.9819212

> SS.yy<-sum((y-mean(y))^2)
>
> SS.xy/sqrt(SS.xx*SS.yy)

+ [1] 0.9819212
```

The Coefficient of Correlation

Note: $-1 \leqslant r \leqslant 1$.

- If r is close to 1, there is a strong linear relationship between x and y where y increases with x.
- If r is close to -1, there is a strong linear relationship between x and y where y decreases with x.

Note: In the model

$$y = \beta_0 + \beta_1 x$$

$\beta_1 = 0 \implies r \approx 0$, so tests for $\beta_1 = 0$ can also be used to deduce a lack of correlation between the variables.

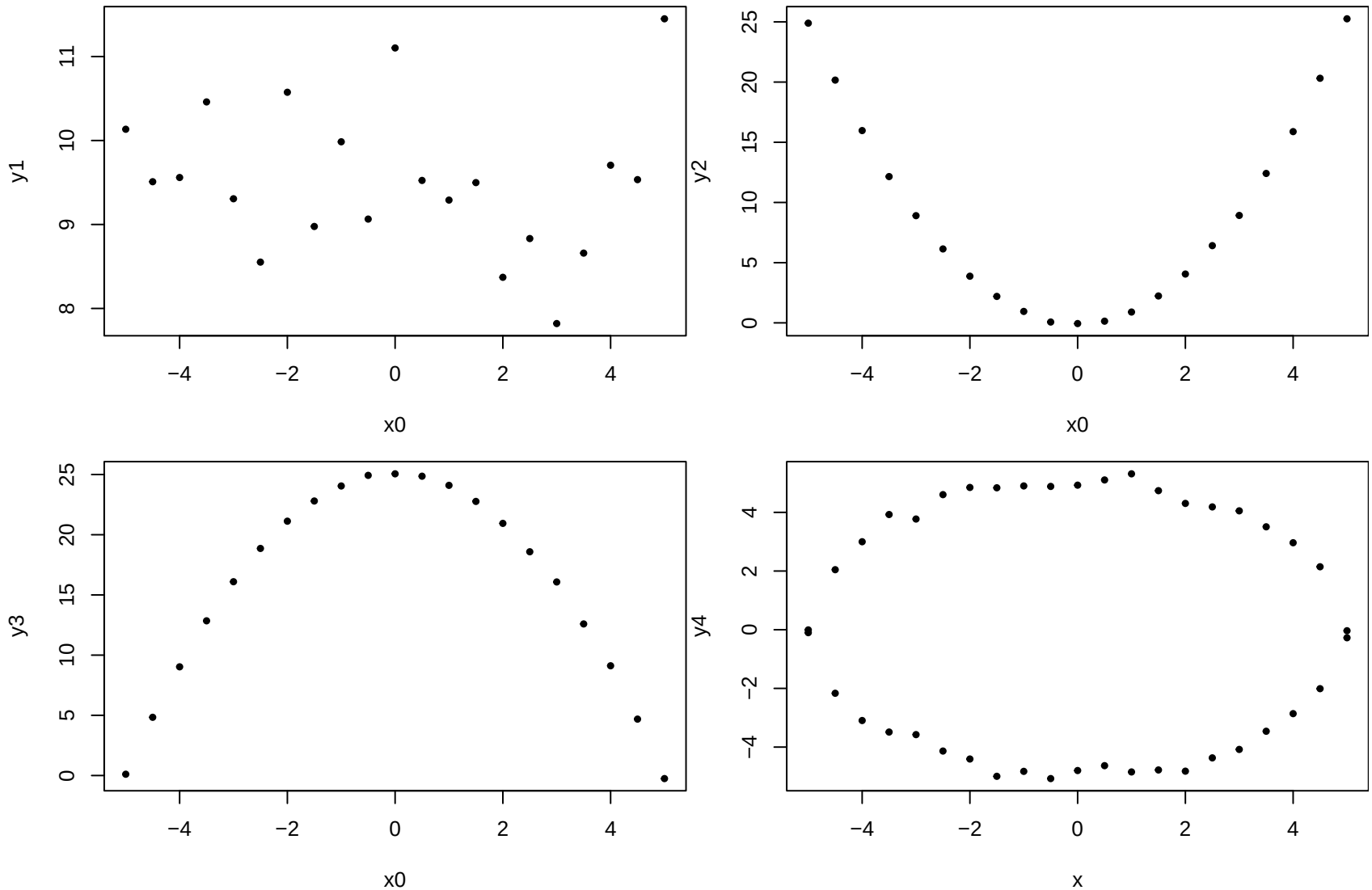
The Coefficient of Correlation

Notes:

1. A strong linear relationship is not necessarily a causal relationship, that is, just because $r \approx 1$ does not mean that x causes changes in y (we may have a spurious correlation).
2. Just because $r \approx 0$ does not mean that that x and y are unrelated, merely that they are uncorrelated.

   That is, it is possible to construct examples where x and y have a strong functional relationship, but where $r = 0$.

The Coefficient of Correlation



The Coefficient of Correlation

All of these data sets have $r \approx 0$.

- top left: no systematic relationship
- top right: almost deterministic parabolic relationship
- bottom left: almost deterministic parabolic relationship
- bottom right: data lie on a circle

Testing Correlation

We use ρ to denote the true correlation between X and Y.

We can test the hypothesis that $\rho = 0$ (that is, that X and Y are uncorrelated) using the statistic r. For testing

$$\begin{aligned} H_0 &: \rho = 0 \\ H_a &: \rho \neq 0 \end{aligned}$$

we can use the test statistic

$$T = \frac{r}{\sqrt{(1 - r^2)/(n - 2)}}$$

If H_0 is true, then approximately

$$T \sim t_{n-2}.$$

Testing Correlation

Alternately, we could use

$$Z = \frac{1}{2} \log \left(\frac{1+r}{1-r} \right)$$

and then, if H_0 is true, we have that approximately

$$Z \sim N \left(\frac{1}{2} \log \left(\frac{1+\rho}{1-\rho} \right), \frac{1}{n-3} \right)$$

when $\rho = 0$, so that (approximately)

$$\sqrt{n-3}\, Z \sim Normal(0,1)$$

R^2 statistic

A related quantity is the

Coefficient of Determination

or R^2 statistic

$$r^2 = \frac{SS_{yy} - SSE}{SS_{yy}} = 1 - \frac{SSE}{SS_{yy}}$$

R^2 statistic

Note that the total variation in y is recorded via

$$SS_{yy} = \sum_{i=1}^{n} (y_i - \overline{y})^2$$

and the random variation is recorded via

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

R^2 statistic

Therefore the variation explained by the linear regression is

$$SSR = SS_{yy} - SSE \qquad \text{as} \qquad SS_{yy} = SSR + SSE$$

Thus

$$r^2 = \frac{SSR}{SS_{yy}} = \frac{\text{Variation explained by Regression}}{\text{Total Variation}}$$

r^2 is a measure of model adequacy, that is, if $r^2 \approx 1$, then the linear model is a good fit.

R^2 statistic

In `lm`, the quantity reported as `Multiple R-squared` is the R^2 statistic.

```
> summary(fit1)


+
+ Call:
+ lm(formula = y ~ x)
+
+ Residuals:
+      Min       1Q   Median       3Q      Max
+ -13.9386  -2.9048  -0.7366   1.7927  11.6499
+
+ Coefficients:
+             Estimate Std. Error t value Pr(>|t|)
+ (Intercept)  -4.8517     3.0097  -1.612    0.124
+ x             5.4887     0.2494  22.008 1.84e-14 ***
+ ---
+ Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
+
+ Residual standard error: 5.674 on 18 degrees of freedom
+ Multiple R-squared:  0.9642, Adjusted R-squared:  0.9622
+ F-statistic: 484.4 on 1 and 18 DF,  p-value: 1.835e-14
```

R^2 statistic

```
> cor(x,y)^2

+ [1] 0.9641693

> summary(fit1)$r.squared

+ [1] 0.9641693
```

R^2 statistic

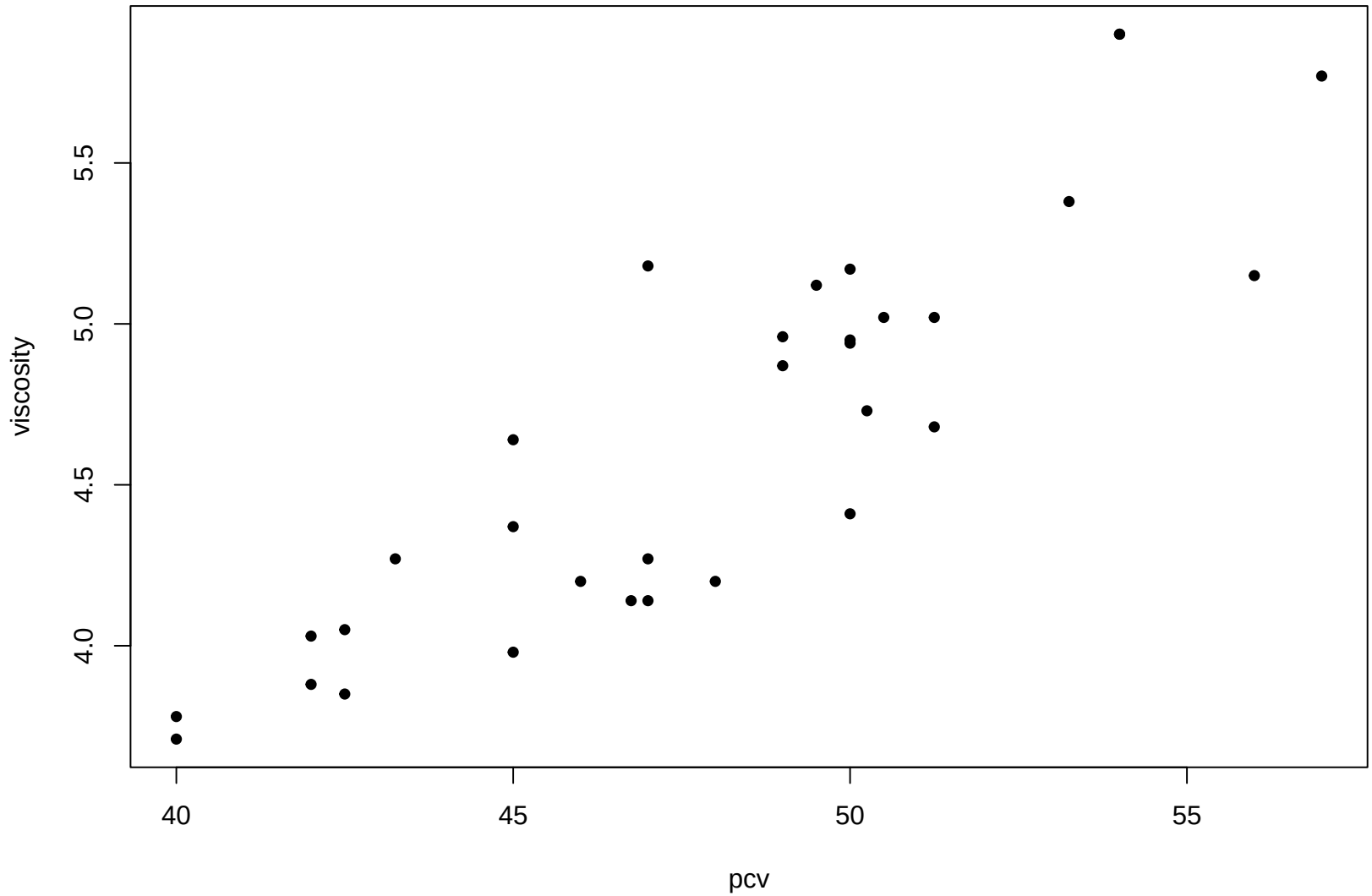
Example: Blood viscosity vs PCV

```
> load('204DataSets.rda')
> head(Viscosity)

+   Patient viscosity  pcv plasfib plaspro
+ 1       1      3.71 40.0     344    6.27
+ 2       2      3.78 40.0     330    4.86
+ 3       3      3.85 42.5     280    5.09
+ 4       4      3.88 42.0     418    6.79
+ 5       5      3.98 45.0     774    6.40
+ 6       6      4.03 42.0     388    5.48

> par(mar=c(4,4,2,2))
> plot(viscosity~pcv,data=Viscosity,pch=19,cex=0.8)
```

R^2 statistic



R^2 statistic

```
> fit.vis<-lm(viscosity~pcv,data=Viscosity)
> summary(fit.vis)

+
+ Call:
+ lm(formula = viscosity ~ pcv, data = Viscosity)
+
+ Residuals:
+      Min       1Q   Median       3Q      Max
+ -0.48814 -0.22152  0.04686  0.12955  0.64915
+
+ Coefficients:
+             Estimate Std. Error t value Pr(>|t|)
+ (Intercept) -1.22336    0.58422  -2.094   0.0448 *
+ pcv          0.12243    0.01214  10.088 3.73e-11 ***
+ ---
+ Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
+
+ Residual standard error: 0.3012 on 30 degrees of freedom
+ Multiple R-squared:  0.7723, Adjusted R-squared:  0.7647
+ F-statistic: 101.8 on 1 and 30 DF,  p-value: 3.731e-11
```

R^2 statistic

```
> cor(Viscosity$pcv,Viscosity$viscosity)

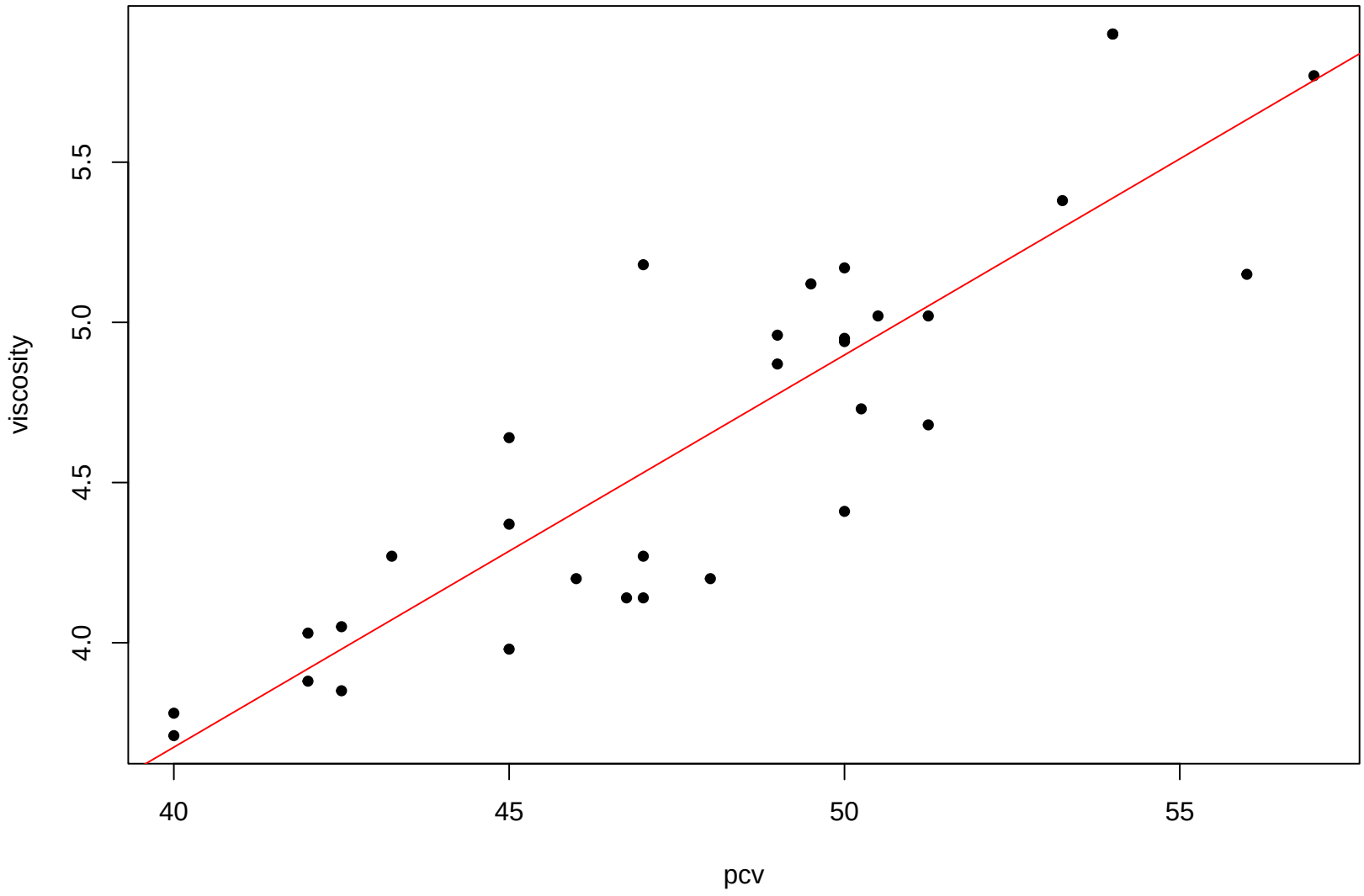
+ [1] 0.878818

> cor(Viscosity$pcv,Viscosity$viscosity)^2

+ [1] 0.7723211

> par(mar=c(4,4,2,2))
> plot(viscosity~pcv,data=Viscosity,pch=19,cex=0.8)
> abline(coef(fit.vis),col='red')
```

R^2 statistic



R^2 statistic

Example: Blood Viscosity vs PCV

We have

- $n = 32$
- $r = 0.879$
- $R^2 = r^2 = (0.879)^2 = 0.772$

Test of $\rho = 0$:

$$T_{obs} = \frac{r}{\sqrt{(1 - r^2)/(n - 2)}} = 10.087$$

We compare with a $t_{n-2} \equiv t_{30}$ distribution; the p-value is 3.73×10^{-11}, so there is strong evidence that $\rho \neq 0$.

Prediction

After the straight-line model is fitted, it can be used for forecasting or prediction. That is, given a new x value we can predict the corresponding y.

At any value x_p, the prediction $\widehat{y}_p$ is

$$\widehat{y}_p = \widehat{\beta}_0 + \widehat{\beta}_1 \mathbf{x}_p$$

This is the best predictor of Y at this x value.

Prediction

We can also compute the standard error of this prediction; if the value of the random error variance σ^2 is known, then

$$\text{s.e.}(\widehat{y}_p) = \sigma\sqrt{\frac{1}{n} + \frac{(x_p - \overline{x})^2}{SS_{xx}}}$$

If σ is unknown, we estimate σ by $\widehat{\sigma} = s$ as defined previously

$$s^2 = \frac{SSE(\widehat{\beta}_0, \widehat{\beta}_1)}{n-2}$$

so that

$$\text{e.s.e.}(\widehat{y}_p) = s\sqrt{\frac{1}{n} + \frac{(x_p - \overline{x})^2}{SS_{xx}}}$$

Prediction

This prediction is the expected value of y at $x = x_p$. That is, we have worked out

$$Var[\widehat{Y}_p|x_p] = Var[\widehat{\beta}_0 + \widehat{\beta}_1 x_p|x_p]$$

to compute the s.e. for $\widehat{Y}_p$.

But we can actually predict an error corrupted version of $\widehat{Y}_p$, $\widehat{Y}_p^\star$ say, where

$$\widehat{Y}_p^\star = \widehat{Y}_p + \epsilon_p$$

where ϵ_p is a new random error.

Prediction

But

$$\begin{aligned} Var[\widehat{Y}_p^\star | x_p] &= Var[\widehat{Y}_p + \epsilon_p | x_p] \\ &= Var[\widehat{Y}_p | x_p] + Var[\epsilon_p | x_p] \\ &= Var[\widehat{Y}_p | x_p] + \sigma^2 \end{aligned}$$

that is, there is an extra piece of variation due to ϵ_p.

Thus

$$\text{e.s.e.}(\widehat{y}_p^\star) = s\sqrt{1 + \frac{1}{n} + \frac{(x_p - \overline{x})^2}{SS_{xx}}} > \text{e.s.e.}(\widehat{y}_p)$$

Prediction Intervals

A $100(1-\alpha)\%$ prediction interval for the mean value at $x = x_p$ is

$$\widehat{y}_p \pm t_{\alpha/2,n-2} s \sqrt{\frac{1}{n} + \frac{(x_p - \overline{x})^2}{SS_{xx}}}$$

whereas for an individual new value (predicted with error) at $x = x_p$ is

$$\widehat{y}_p \pm t_{\alpha/2,n-2} s \sqrt{1 + \frac{1}{n} + \frac{(x_p - \overline{x})^2}{SS_{xx}}}$$

Prediction Intervals

Using the predict function:

```
> xp<-seq(0,20,by=0.01)
> int.mean<-predict(fit1,newdata=data.frame(x=xp),interval='confidence', se.fit=TRUE)
> int.obs<-predict(fit1,newdata=data.frame(x=xp),interval='prediction', se.fit=TRUE)
> names(int.mean)

+ [1] "fit"            "se.fit"         "df"             "residual.scale"

> head(int.mean$fit)

+         fit       lwr      upr
+ 1 -4.851720 -11.17481 1.471369
+ 2 -4.796834 -11.11517 1.521504
+ 3 -4.741947 -11.05554 1.571641
+ 4 -4.687060 -10.99590 1.621778
+ 5 -4.632174 -10.93626 1.671916
+ 6 -4.577287 -10.87663 1.722055

> head(int.mean$se.fit)

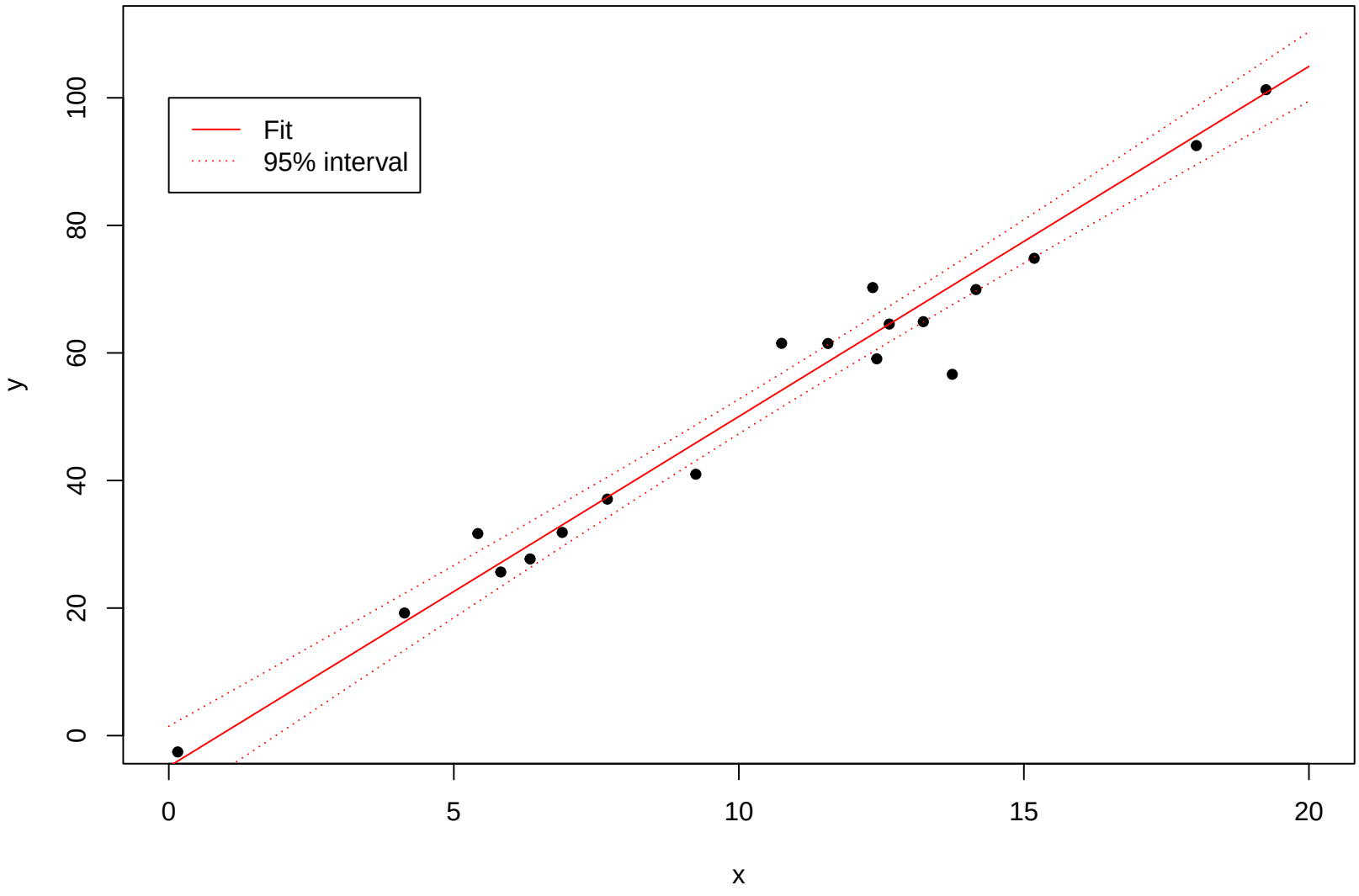
+        1        2        3        4        5        6
+ 3.009673 3.007412 3.005151 3.002890 3.000630 2.998370
```

Prediction Intervals

Prediction interval for mean:

```
> par(mar=c(4,4,2,2))
> plot(x,y,pch=19,cex=0.8,xlim=range(0,20), ylim=range(0,110))
> lines(xp,int.mean$fit[,1],col='red')
> lines(xp,int.mean$fit[,2],col='red',lty=3)    #Lower boundary of interval
> lines(xp,int.mean$fit[,3],col='red',lty=3)    #Upper boundary of interval
> legend(0,100,c('Fit','95% interval'),col='red',lty=c(1,3))
```

Prediction Intervals

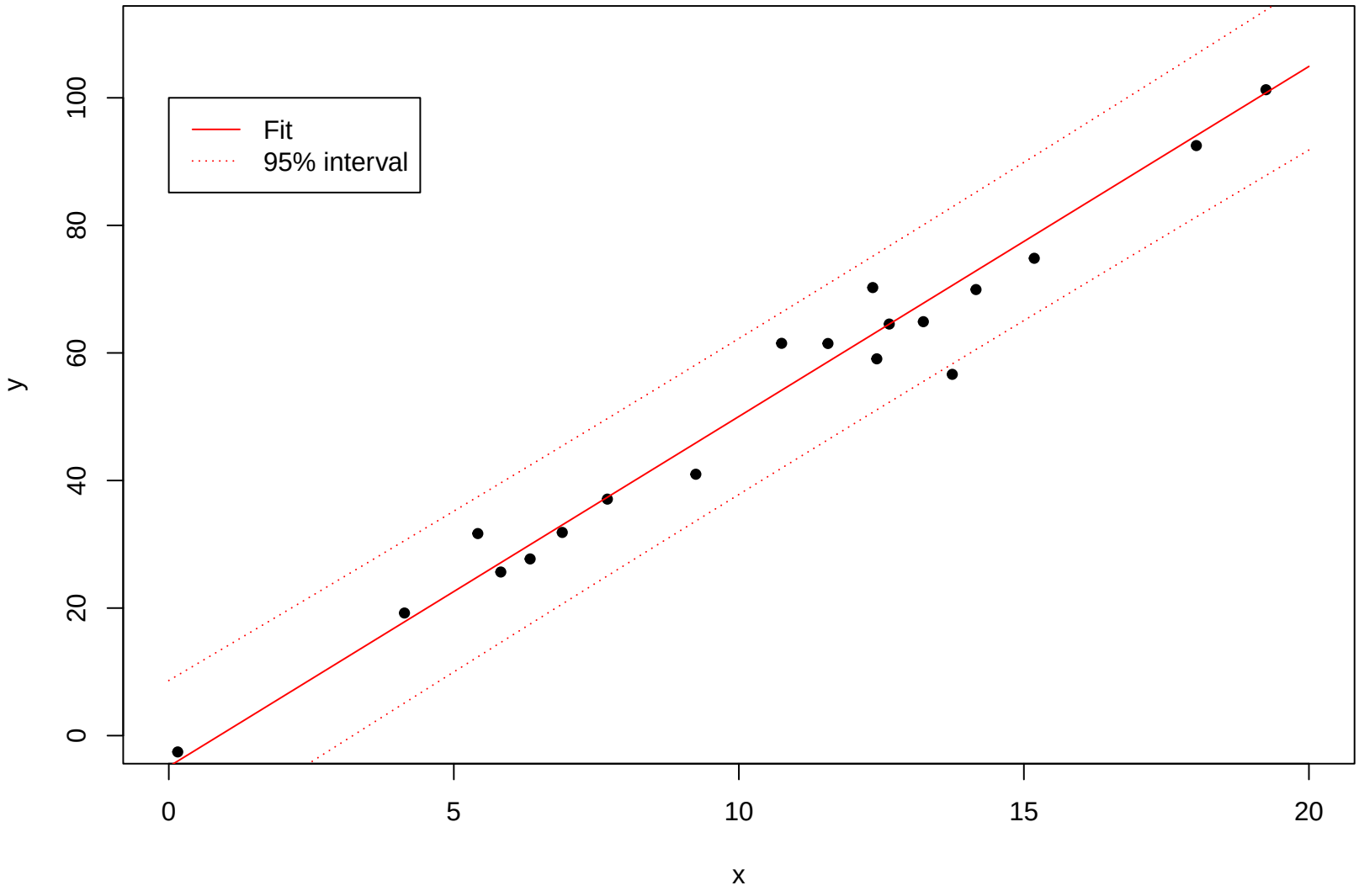


Prediction Intervals

Prediction interval for individual observation:

```
> par(mar=c(4,4,2,2))
> plot(x,y,pch=19,cex=0.8,xlim=range(0,20), ylim=range(0,110))
> lines(xp,int.obs$fit[,1],col='red')
> lines(xp,int.obs$fit[,2],col='red',lty=3)    #Lower boundary of interval
> lines(xp,int.obs$fit[,3],col='red',lty=3)    #Upper boundary of interval
> legend(0,100,c('Fit','95% interval'),col='red',lty=c(1,3))
```

Prediction Intervals



ANOVA F-test in Regression

An ANOVA F-test can be constructed to test overall (global) fit of the linear regression model.

The decomposition of sums of squares for regression takes the form

$$SS = SSR + SSE$$

where

- $SS = SS_{yy}$: overall or total sum of squares
- SSR: sum of squares due to Regression
- SSE: sum of squares due to Error

ANOVA F-test in Regression

$$
\begin{aligned}
SS &= \sum_{i=1}^{n} (y_i - \overline{y})^2 \\
SSR &= \sum_{i=1}^{n} (\widehat{y}_i - \overline{y})^2 \\
SSE &= \sum_{i=1}^{n} (y_i - \widehat{y}_i)^2
\end{aligned}
$$

where

$$\widehat{y}_i = \widehat{\beta}_0 + \widehat{\beta}_1 x_i \qquad i = 1, \ldots, n$$

ANOVA F-test in Regression

Degrees of Freedom

- TOTAL: $n - 1$
- REGRESSION: 1
  - there is one parameter that explains the influence of x on y: β_1.
- ERROR: $n - 2$
  - error d.f. is $n - p$
  - here $p = 2$ records the total number of parameters in the mean model, that is, β_0 and β_1.

The ANOVA Table

SOURCE	DF	SS	MS	F
REGRESSION	1	SSR	$MSR = \frac{SSR}{1}$	$F = \frac{MSR}{MSE}$
ERROR	$n-2$	SSE	$MSE = \frac{SSE}{(n-2)}$	
TOTAL	$n-1$	SS		

The ANOVA Table

The test of the hypothesis

$$\begin{aligned} H_0 &: \mathbb{E}[Y|\mathbf{x}] = \beta_0 \\ H_a &: \mathbb{E}[Y|\mathbf{x}] = \beta_0 + \beta_1 \mathbf{x} \end{aligned}$$

can be completed by using the test statistic

$$F = \frac{MSR}{MSE}$$

If H_0 is true

$$F \sim F_{1,n-2}.$$

The ANOVA Table

```
> summary(aov(lm(y~x)))

+             Df Sum Sq Mean Sq F value   Pr(>F)
+ x            1  15592   15592   484.4 1.84e-14 ***
+ Residuals   18    579      32
+ ---
+ Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

> library(car)
> Anova(lm(y~x))

+ Anova Table (Type II tests)
+
+ Response: y
+            Sum Sq Df F value    Pr(>F)
+ x         15591.8  1  484.36 1.835e-14 ***
+ Residuals   579.4 18
+ ---
+ Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

The ANOVA Table

```
> library(supernova)
> supernova(lm(y~x))

+  Analysis of Variance Table (Type III SS)
+  Model: y ~ x
+
+                                SS df        MS       F   PRE     p
+  ----- --------------- | --------- -- --------- ------- ----- -----
+  Model (error reduced) | 15591.832  1 15591.832 484.363 .9642 .0000
+  Error (from model)    |   579.427 18    32.190
+  ----- --------------- | --------- -- --------- ------- ----- -----
+  Total (empty model)   | 16171.259 19   851.119

> supernova(lm(y~x), type=2)

+  Analysis of Variance Table (Type II SS)
+  Model: y ~ x
+
+                                SS df        MS       F   PRE     p
+  ----- --------------- | --------- -- --------- ------- ----- -----
+  Model (error reduced) | 15591.832  1 15591.832 484.363 .9642 .0000
+  Error (from model)    |   579.427 18    32.190
+  ----- --------------- | --------- -- --------- ------- ----- -----
+  Total (empty model)   | 16171.259 19   851.119
```

The ANOVA Table

This is just like the ANOVA in the one-way layout (CRD) with n groups, but where

$$\mu_i = \beta_0 + \beta_1 \mathbf{x}_i$$

That is, the group means are structured, that is, we have a formula relating the μ_i quantities.

The ANOVA Table

Consider $r = 3$ replicates at x each of the values $(1, 2, 3, 4)$ in a regression, and say

$$Y_{ij} = 2 + x_i + \epsilon_{ij}$$

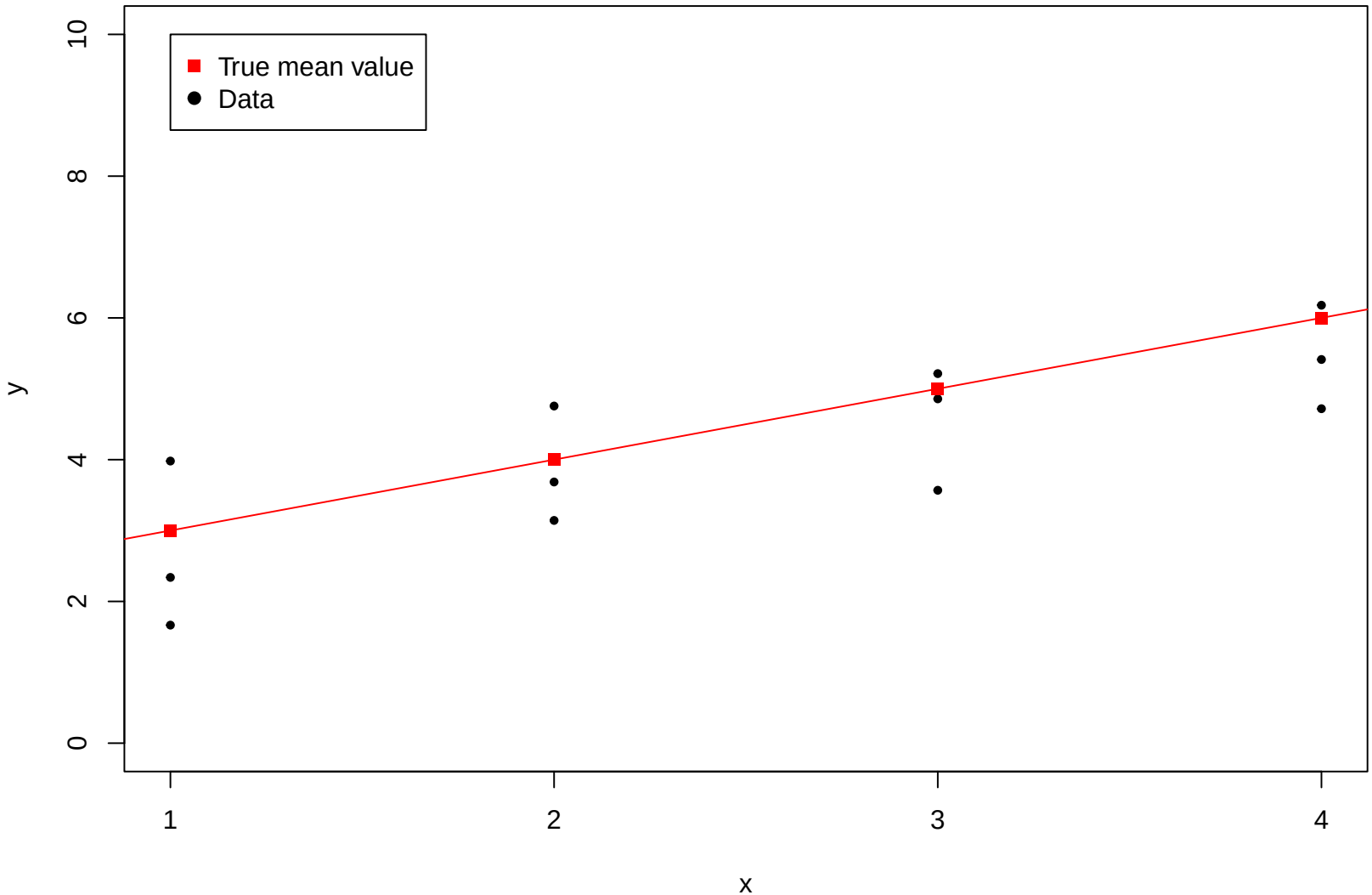
for $i = 1, 2, 3, 4$ and $j = 1, 2, 3$. Then the group means at

$$x_1 = 1 \qquad x_2 = 2 \qquad x_3 = 3 \qquad x_4 = 4$$

are

$$\mu_1 = 2 + x_1 = 3 \quad \mu_2 = 2 + x_2 = 4 \quad \mu_3 = 2 + x_3 = 5 \quad \mu_4 = 2 + x_4 = 6.$$

The ANOVA Table



The ANOVA Table

In a standard one-way layout, we would have four means that are treated as separate parameters

$$\mu_1, \mu_2, \mu_3, \mu_4$$

but in the regression case we have the formula

$$\mu_i = \beta_0 + \beta_1 \mathbf{x}_i$$

so the means are specified using only two parameters (β_0, β_1).

The ANOVA Table

Blood viscosity data:

```
> fit.vis<-lm(viscosity~pcv, data=Viscosity)
> summary(fit.vis)


+
+ Call:
+ lm(formula = viscosity ~ pcv, data = Viscosity)
+
+ Residuals:
+      Min       1Q   Median       3Q      Max
+ -0.48814 -0.22152  0.04686  0.12955  0.64915
+
+ Coefficients:
+             Estimate Std. Error t value Pr(>|t|)
+ (Intercept) -1.22336    0.58422  -2.094   0.0448 *
+ pcv          0.12243    0.01214  10.088 3.73e-11 ***
+ ---
+ Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
+
+ Residual standard error: 0.3012 on 30 degrees of freedom
+ Multiple R-squared:  0.7723, Adjusted R-squared:  0.7647
+ F-statistic: 101.8 on 1 and 30 DF,  p-value: 3.731e-11
```

The ANOVA Table

```
> summary(aov(fit.vis))

+             Df Sum Sq Mean Sq F value   Pr(>F)
+ pcv          1  9.230   9.230   101.8 3.73e-11 ***
+ Residuals   30  2.721   0.091
+ ---
+ Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

Note that the p-values in the

- t-test of the slope coefficient
- ANOVA test

are identical. This is because the tests are equivalent, and

$$T_{obs}^2 = F_{obs}.$$

Checking the Local Fit

A plot of the residuals

$$\hat{e}_i = y_i - \hat{y}_i$$

can reveal model inadequacies. We should observe that in plots of

- $\hat{e}$ vertical vs x horizontal
- $\hat{e}$ vertical vs $\hat{y}$ horizontal

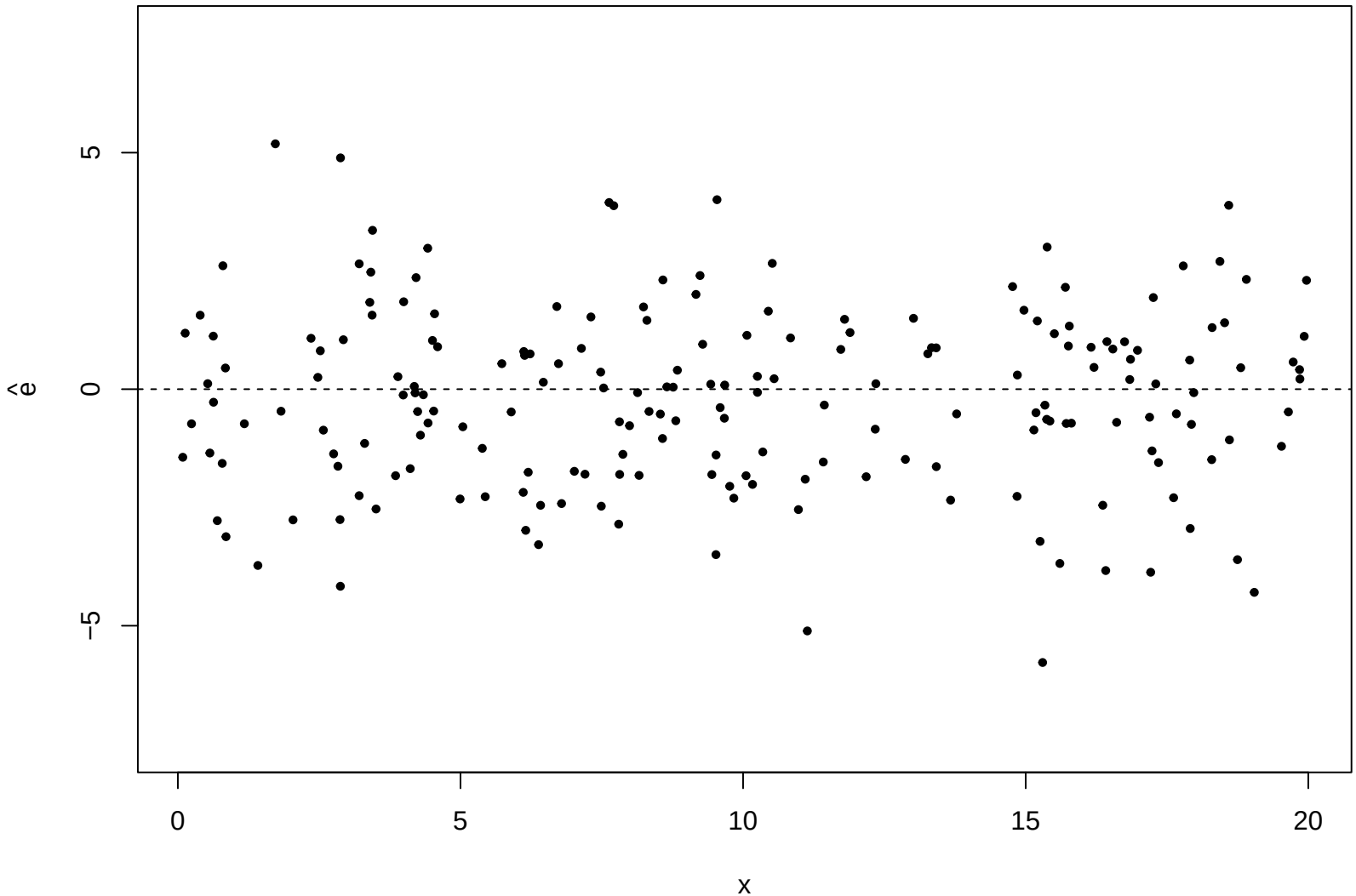
there is no discernible pattern

Checking the Local Fit

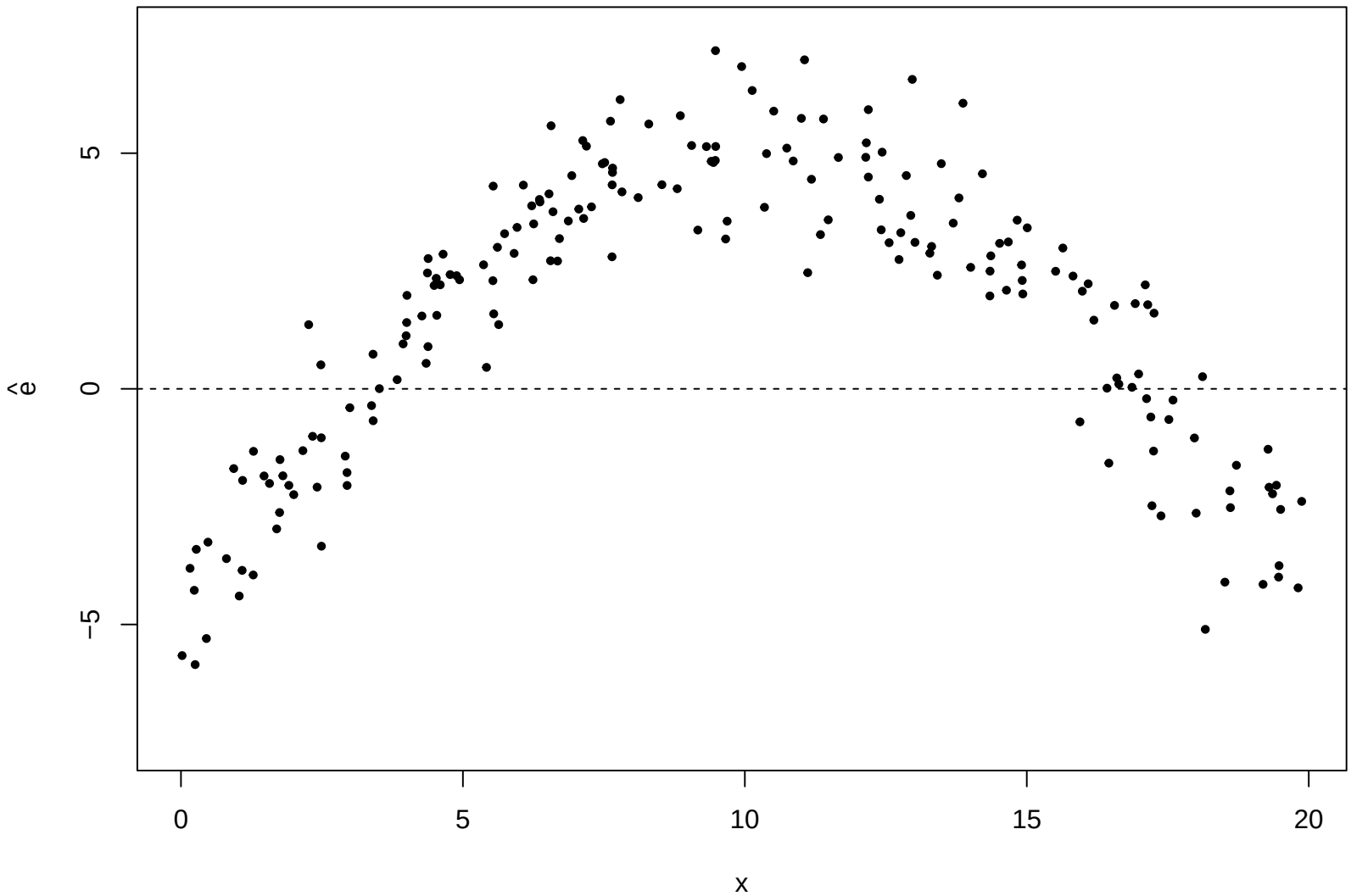
Specifically these plots should

- be vertically centered at zero for every horizontal value
- have the same variability around zero for every horizontal value

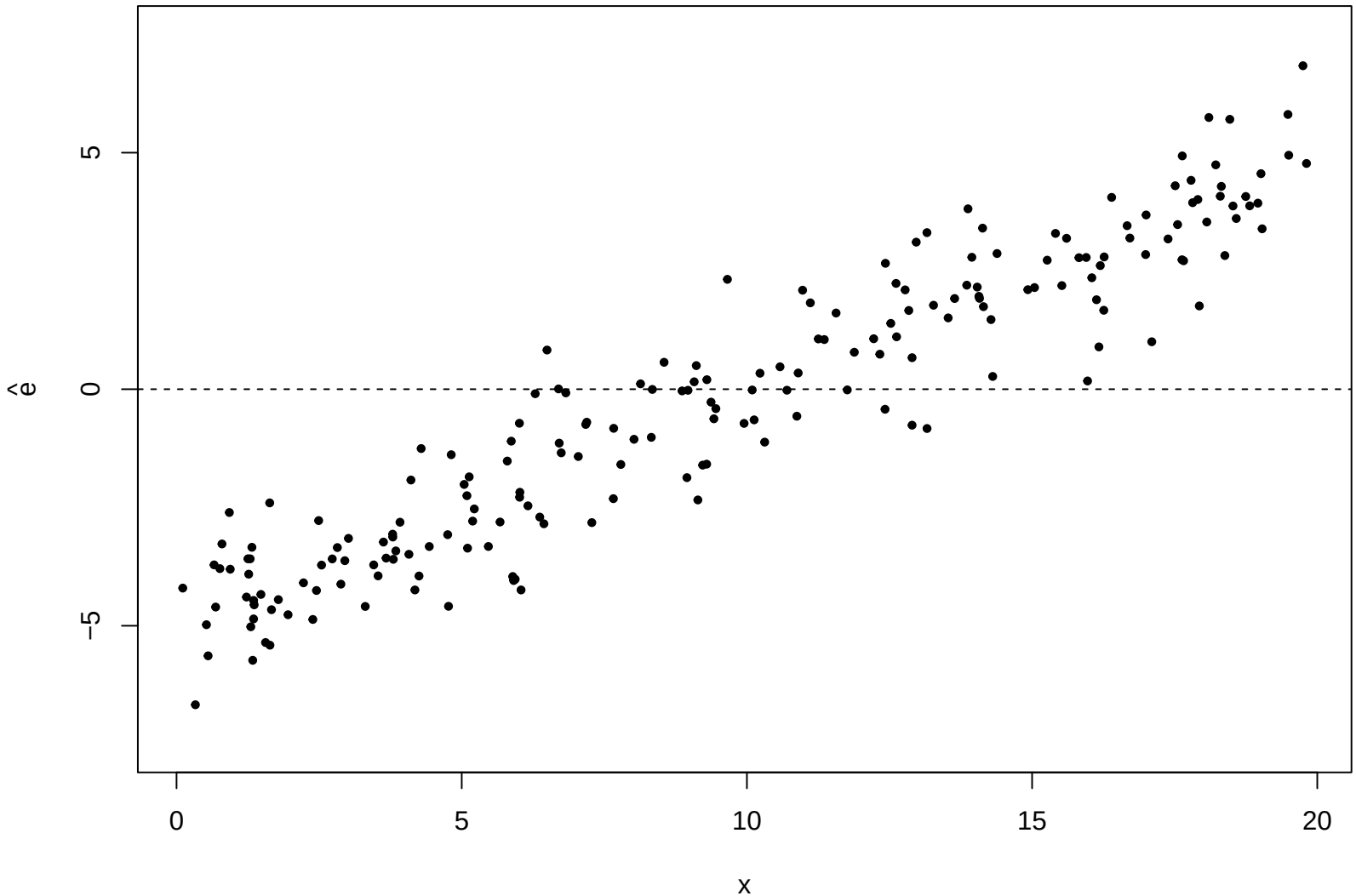
Good fit: no residual pattern



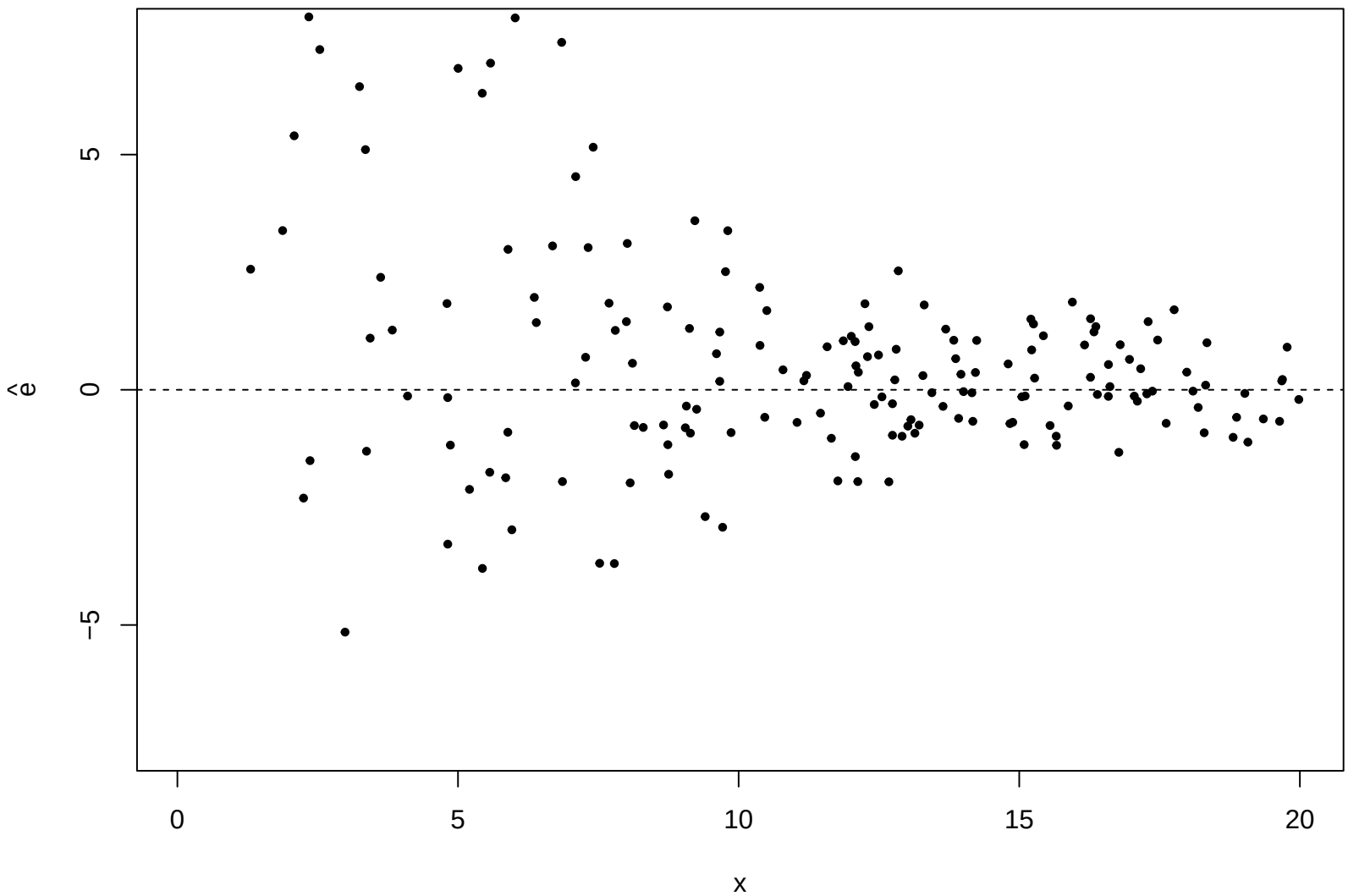
Poor fit: $\hat{e}$ not mean zero for all x



Poor fit: $\hat{e}$ not mean zero for all x



Poor fit: $\hat{e}$ not constant variance



Residual plots: notes

(i) It is rarely possible to make a definitive conclusion from inspecting a plot

(ii) A residual plot with which does not have mean zero for $\hat{e}$ for all values indicates that the proposed regression model is likely incorrect. That is, we assumed that

$$\mathbb{E}[Y|x] = \beta_0 + \beta_1 x$$

but in fact

$$\mathbb{E}[Y|x] = \beta_0 + \beta_1 x + h(x)$$

is the true (data generating) model, for some $h(x)$.

R^2 and adjusted R^2

Model assessment can be made using either the R^2 statistic

$$R^2 = 1 - \frac{SSE}{SS}$$

and the adjusted R^2 statistic

$$R^2_{adj} = 1 - \frac{SSE/EDF}{SS/TDF}$$

where

- EDF = error degrees of freedom = $n - 2$
- TDF = total degrees of freedom = $n - 1$

Polynomial Regression

In many practical situations, the simple straight-line

$$y = \beta_0 + \beta_1 x$$

is not appropriate. Instead, a model including powers of x

$$x^2, x^3, \ldots, x^k$$

should be considered. For example

$$y = \beta_0 + \sum_{j=1}^{k} \beta_j x^j = \beta_0 + \beta_1 x + \cdots + \beta_k x^k$$

Polynomial Regression

The Polynomial Regression Model

$$Y = \beta_0 + \beta_1 \mathbf{x} + \cdots + \beta_k \mathbf{x}^k + \epsilon$$

where ϵ is a random error term as before can be used to model data.

Polynomial Regression

Two immediate questions:

1. How to carry out inference ?
    - estimation
    - testing
    - prediction
2. How to choose k ?

We begin by addressing 1. The estimation of parameters can be again carried out using the principle of least squares provided that the model assumptions listed before are valid.

$k = 2$: quadratic regression

We choose the three parameters $\beta = (\beta_0, \beta_1, \beta_2)^\top$ to minimize the sum of squared errors

$$SSE(\beta) = \sum_{i=1}^{n} (y_i - \hat{y}_i)^2 = \sum_{i=1}^{n} (y_i - \beta_0 - \beta_1 x_i - \beta_2 x_i^2)^2$$

that is the fitted values for parameters β are

$$\hat{y}_i = \beta_0 + \beta_1 x_i + \beta_2 x_i^2$$

$k = 2$: quadratic regression

The estimates $\widehat{\beta}$ can be found to minimize SSE using calculus techniques (differentiating with respect to the elements of β) to give the minimum SSE

$$SSE(\beta) = \sum_{i=1}^{n} (y_i - \widehat{\beta}_0 - \widehat{\beta}_1 x_i - \widehat{\beta}_2 x_i^2)^2$$

$k = 2$: quadratic regression

We can also compute

- the estimated standard errors, which allow tests of parameters to be carried out, and confidence intervals calculated.

- prediction intervals.

- the estimate of the residual error variance σ^2, which is

$$\widehat{\sigma}^2 = \frac{SSE(\widehat{\beta})}{n-p}$$

  Here p is the number of parameters estimated, three, so we divide by $n - 3$.

$k = 2$: quadratic regression

- residuals
  - can be used to assess the fit of the model.
  - the residuals should be patternless if the model fit is good.

- R^2 and Adjusted R^2 statistics
  - used to assess the global fit of the model.
  - used to compare the quality of fit with other models.

Example: Pressure data

Example: Hooker's Pressure Data

This is a standard data set in R: from the help file

> *Forbes, J. (1957). Further experiments and remarks on the measurement of heights by boiling point of water. Transactions of the Royal Society of Edinburgh,* ***21****, 235-243.*
>
> *In his original paper, Forbes provided additional data collected by the botanist Dr. Joseph Hooker on temperatures and boiling points measured often at higher altitudes in the Himalaya Mountains.*

Example: Pressure data

Example: Hooker's Pressure Data

For the Hooker pressure data, a quadratic polynomial ($k = 2$) might be suitable.

$$Y = \beta_0 + \beta_1 x + \beta_2 x^2 + \epsilon$$

We need to estimate β_0, β_1 and β_2 for these data to see if the model fits better than the straight line model we fitted previously.

It transpires that the quadratic model produces a set of residuals that are patternless, which the straight line model when fitted does not.

Example: Pressure data

In the data set there are $n = 31$ observations.

- `pressure` is the x variable, representing the measured air pressure, measured in inches of Mercury
- `temp` is the y variable, representing the measured boiling temperature, measured in degrees F.

```
> head(Hooker)

+    temp pressure
+ 1 210.8   29.211
+ 2 210.2   28.559
+ 3 208.4   27.972
+ 4 202.5   24.697
+ 5 200.6   23.726
+ 6 200.1   23.369

> nrow(Hooker)

+ [1] 31
```

Example: Pressure data

```
> summary(lm(temp~pressure+I(pressure^2), data=Hooker))

+
+ Call:
+ lm(formula = temp ~ pressure + I(pressure^2), data = Hooker)
+
+ Residuals:
+      Min       1Q   Median       3Q      Max
+ -0.79906 -0.26314 -0.01578  0.25139  0.73891
+
+ Coefficients:
+                 Estimate Std. Error t value Pr(>|t|)
+ (Intercept)   126.701623   2.112363  59.981  < 2e-16 ***
+ pressure        4.157627   0.199069  20.885  < 2e-16 ***
+ I(pressure^2)  -0.043754   0.004552  -9.612 2.29e-10 ***
+ ---
+ Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
+
+ Residual standard error: 0.3956 on 28 degrees of freedom
+ Multiple R-squared:  0.9981, Adjusted R-squared:  0.998
+ F-statistic:  7307 on 2 and 28 DF,  p-value: < 2.2e-16
```

We need to use `I(pressure^2)` to compute the quadratic term in the formula.

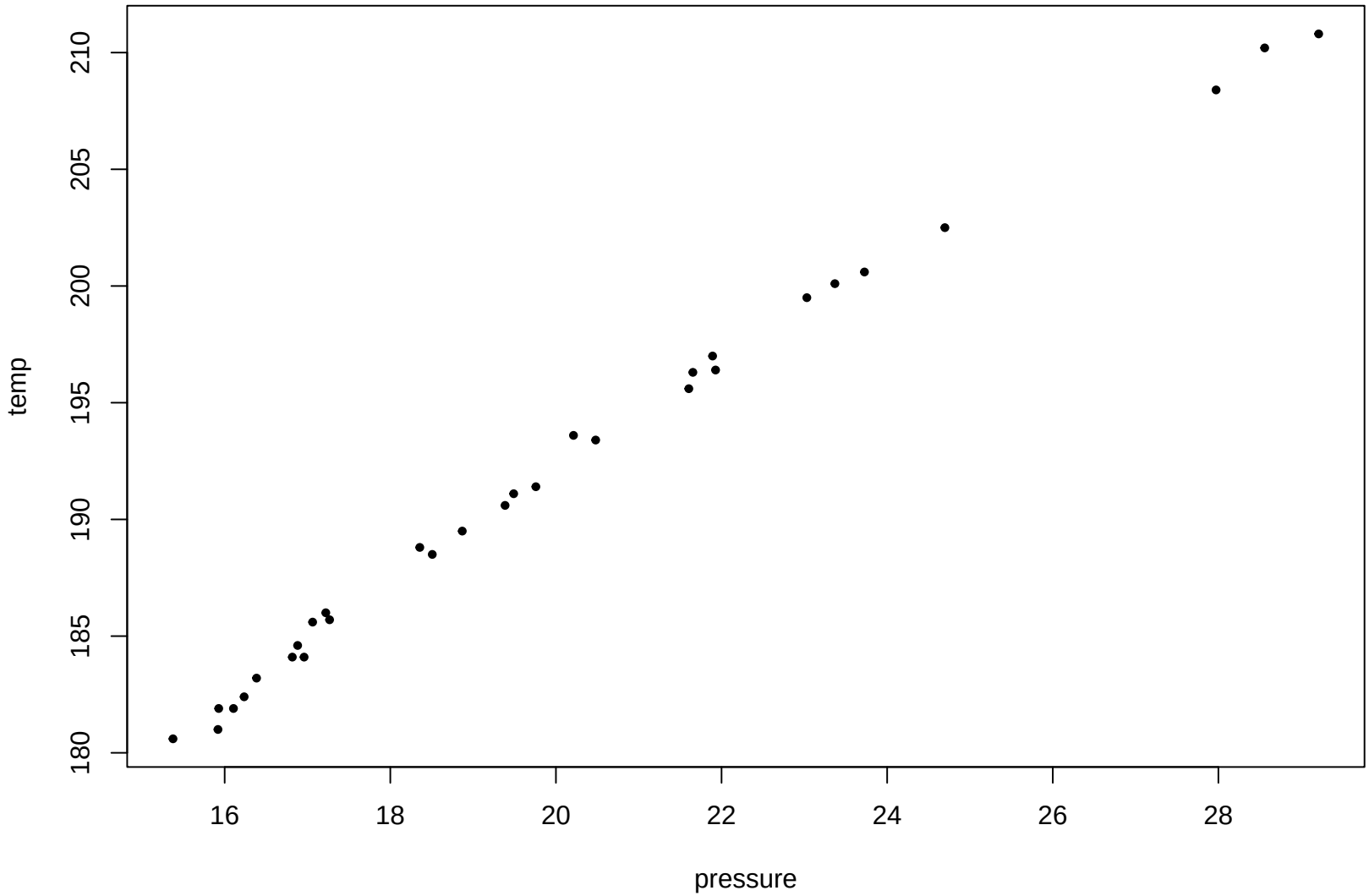
Example: Pressure data

We might also consider a straight-line fit:

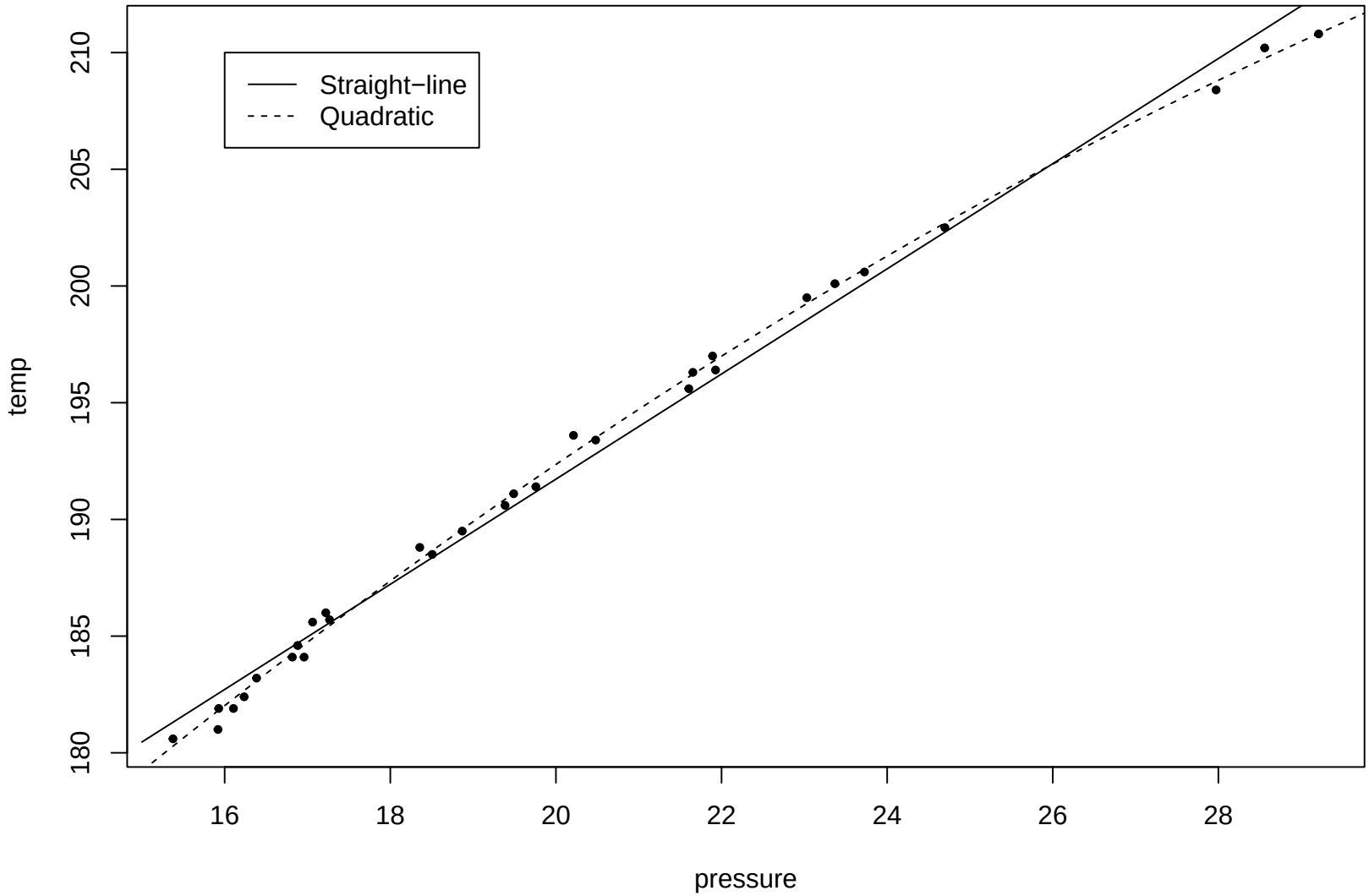
```
> summary(lm(temp~pressure, data=Hooker))

+
+ Call:
+ lm(formula = temp ~ pressure, data = Hooker)
+
+ Residuals:
+     Min      1Q  Median      3Q     Max
+ -1.6735 -0.6805  0.2203  0.5296  1.3976
+
+ Coefficients:
+              Estimate Std. Error t value Pr(>|t|)
+ (Intercept) 146.67290    0.77641  188.91   <2e-16 ***
+ pressure      2.25260    0.03809   59.14   <2e-16 ***
+ ---
+ Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
+
+ Residual standard error: 0.806 on 29 degrees of freedom
+ Multiple R-squared:  0.9918, Adjusted R-squared:  0.9915
+ F-statistic:  3498 on 1 and 29 DF,  p-value: < 2.2e-16
```

Example: Pressure data



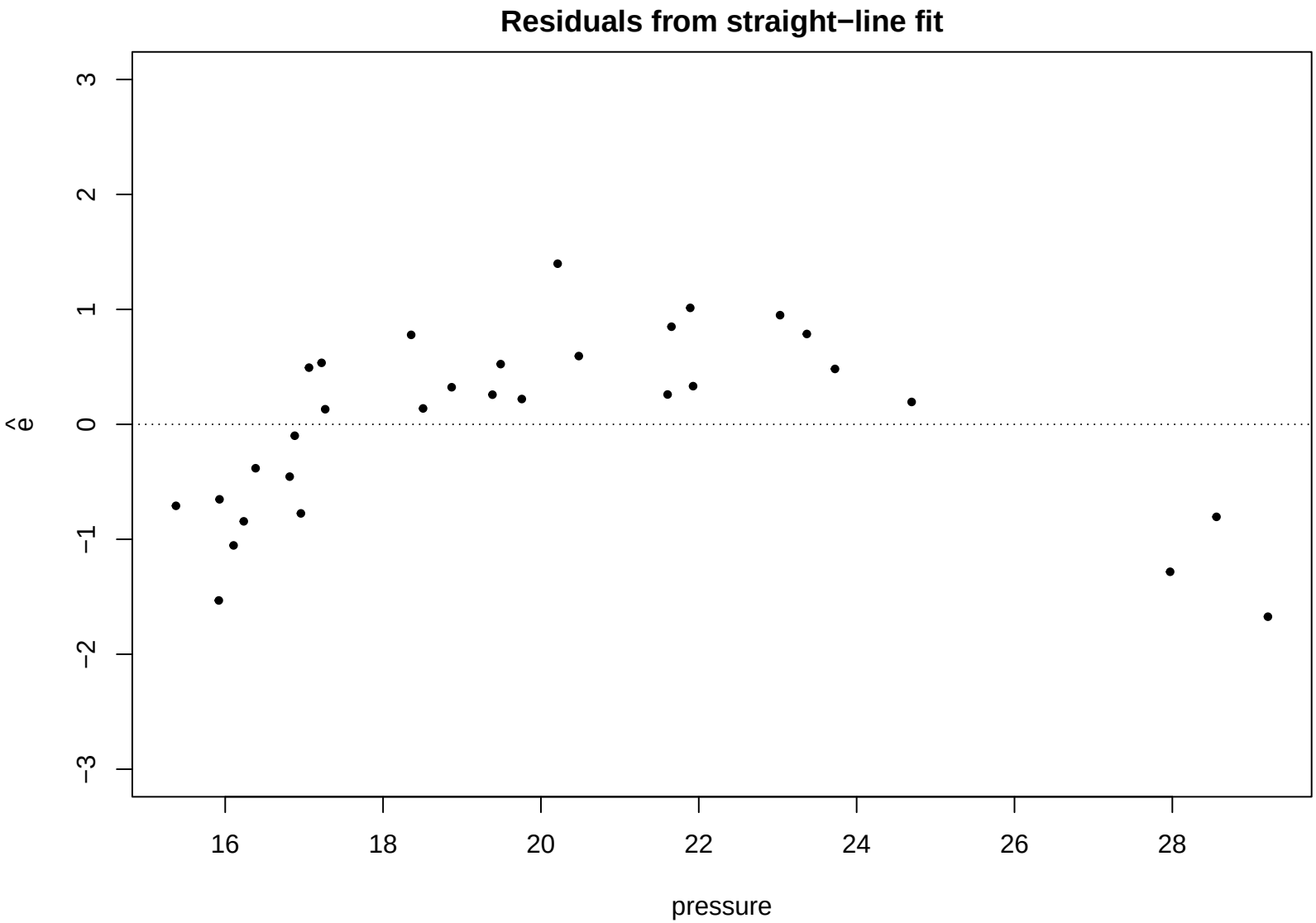
Example: Pressure data



Example: Pressure data

```
> fit.lin<-lm(temp~pressure, data=Hooker)
> fit.quad<-lm(temp~pressure+I(pressure^2), data=Hooker)
> ehat.lin<-residuals(fit.lin)
> ehat.quad<-residuals(fit.quad)
> par(mar=c(4,5,2,2))
> plot(Hooker$pressure, ehat.lin, ylim=range(-3,3),
+         pch=19, cex=0.6, xlab='pressure', ylab=expression(hat(e)))
> abline(h=0,lty=3)
> title('Residuals from straight-line fit')
```

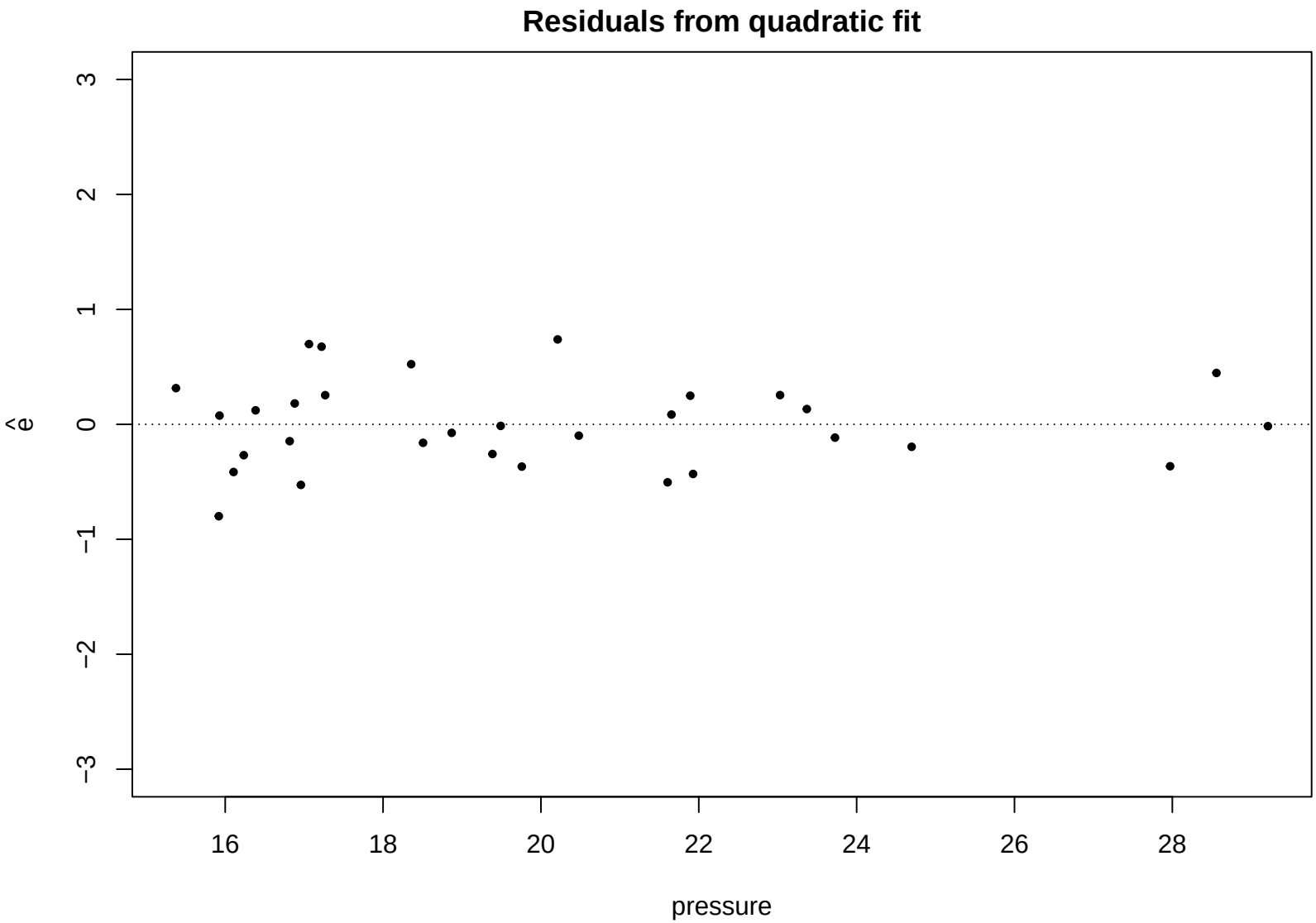
Example: Pressure data



Example: Pressure data

```
> par(mar=c(4,5,2,2))
> plot(Hooker$pressure, ehat.quad, ylim=range(-3,3),
+         pch=19, cex=0.6, xlab='pressure', ylab=expression(hat(e)))
> abline(h=0,lty=3)
> title('Residuals from quadratic fit')
```

Example: Pressure data



Example: Pressure data

It is common to use the standardized residuals

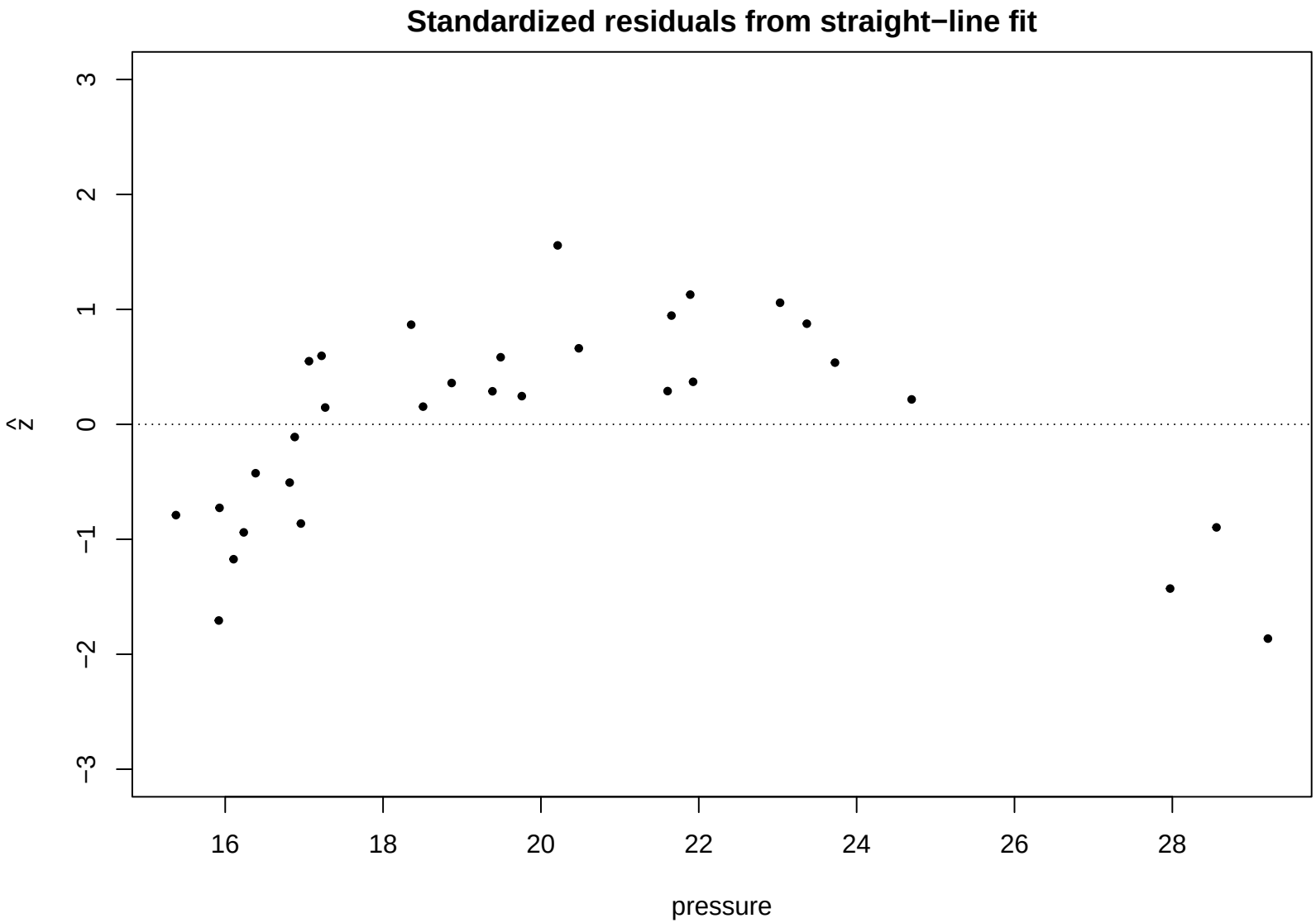
$$\hat{z}_i = \frac{\hat{e}_i}{\hat{\sigma}} = \frac{y_i - \hat{y}_i}{\hat{\sigma}}$$

where $\hat{\sigma}^2$ is the estimate of σ^2 defined previously, as

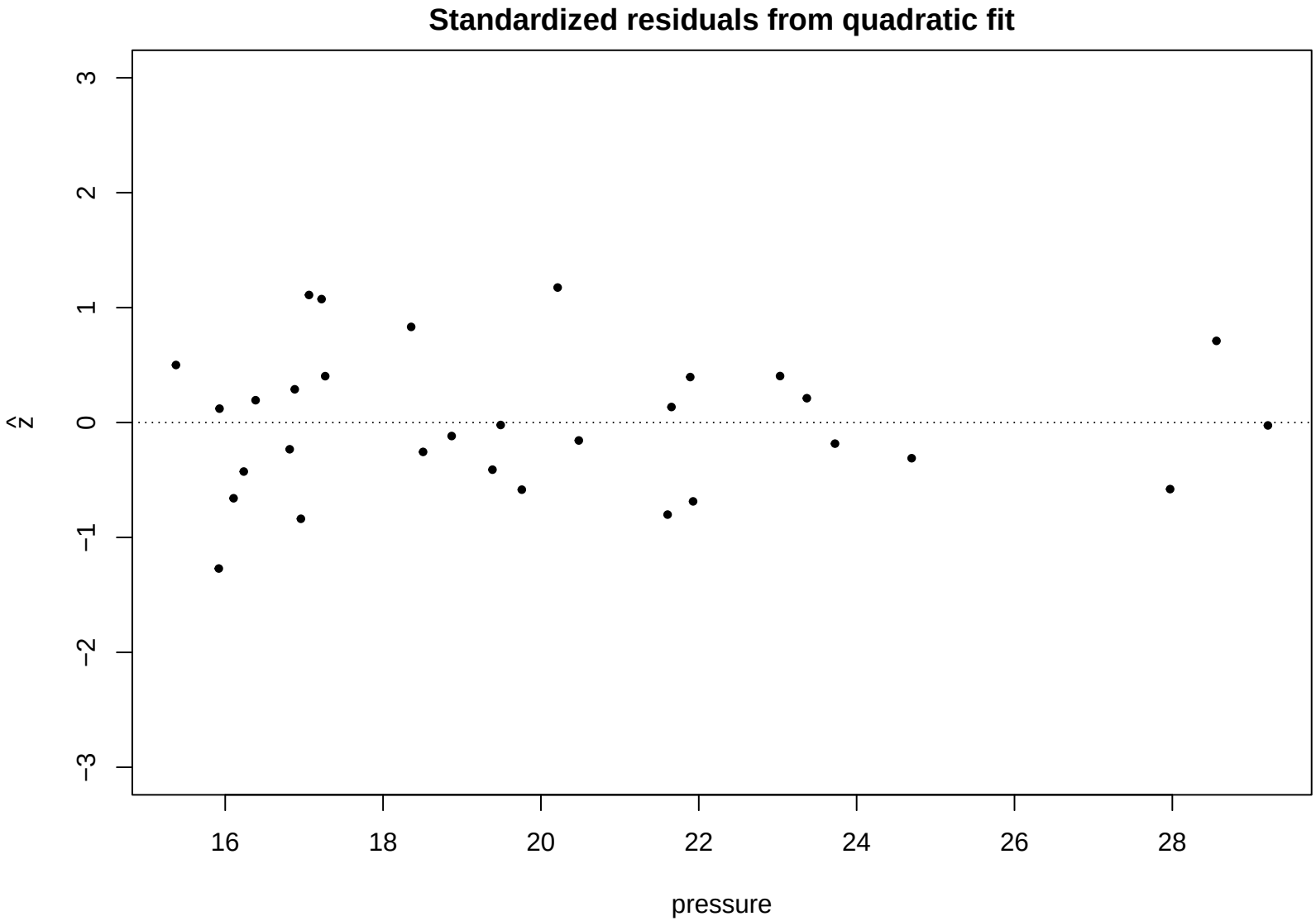
$$\text{Var}[\hat{z}_i] \approx 1$$

It is easier to compare $\hat{e}_i$ across different models when inspecting residuals.

Example: Pressure data



Example: Pressure data



Example: Pressure data

Note:

```
> summary(fit.lin)$coef   #Straight-line fit estimates

+               Estimate Std. Error   t value     Pr(>|t|)
+ (Intercept) 146.672897 0.77641329 188.91085 2.266565e-46
+ pressure      2.252596 0.03808725  59.14306 8.634599e-32

> summary(fit.quad)$coef  #Quadratic fit estimates

+                   Estimate  Std. Error   t value     Pr(>|t|)
+ (Intercept)   126.70162290 2.112363414 59.980978 4.024869e-31
+ pressure        4.15762721 0.199069236 20.885333 1.298517e-18
+ I(pressure^2)  -0.04375373 0.004551914 -9.612162 2.287907e-10
```

The coefficients of `pressure` are appreciably different in the two models.

Example: Pressure data

Even though we describe the two models using the same parameter symbols

$$\begin{aligned} \text{Straight-line} \quad \mathbb{E}[Y|x] &= \beta_0 + \beta_1 x \\ \text{Quadratic} \quad \mathbb{E}[Y|x] &= \beta_0 + \beta_1 x + \beta_2 x^2 \end{aligned}$$

in fact, each term can be only be interpreted within the model it sits.

The Linear Model

Note: Although the model based on

$$y = \beta_0 + \beta_1 x + \beta_2 x^2$$

is not linear in x, it is linear in the parameters. Because of this, we term this a linear model.

It is this fact that makes the least-squares solutions easy to find.

The Linear Model

This model is no more difficult to fit than the model

$$y = \beta_0 + \beta_1 \frac{x}{1+x} + \beta_2(1 - e^{-x})$$

say - it is still a linear in the parameters model. It is in the general class of models

$$y = \beta_0 + \beta_1 g_1(x) + \beta_2 g_2(x)$$

where $g_1(x)$ and $g_2(x)$ are general functions of x.

The Linear Model

In fact, any model of the form

$$y = \sum_{j=0}^{k} \beta_j g_j(\mathbf{x}) + \epsilon$$

can be fitted routinely using least-squares; if we know x, then we can simply compute

$$g_0(\mathbf{x}), g_1(\mathbf{x}), \ldots, g_k(\mathbf{x})$$

and plug those values into the formula..

The Linear Model

Example: Harmonic regression: gene expression data

Measure the gene expression for 43 genes across a 48 hour cycle.

- gene expression measured on the log scale
- data are standardized to have mean zero.

Example: Harmonic regression: gene expression data

Let

$$
\begin{aligned}
g_0(x) &= 1 \\
g_j(x) &= \begin{cases} \cos(\lambda_j x) & j \text{ odd} \\ \sin(\lambda_j x) & j \text{ even} \end{cases}
\end{aligned}
$$

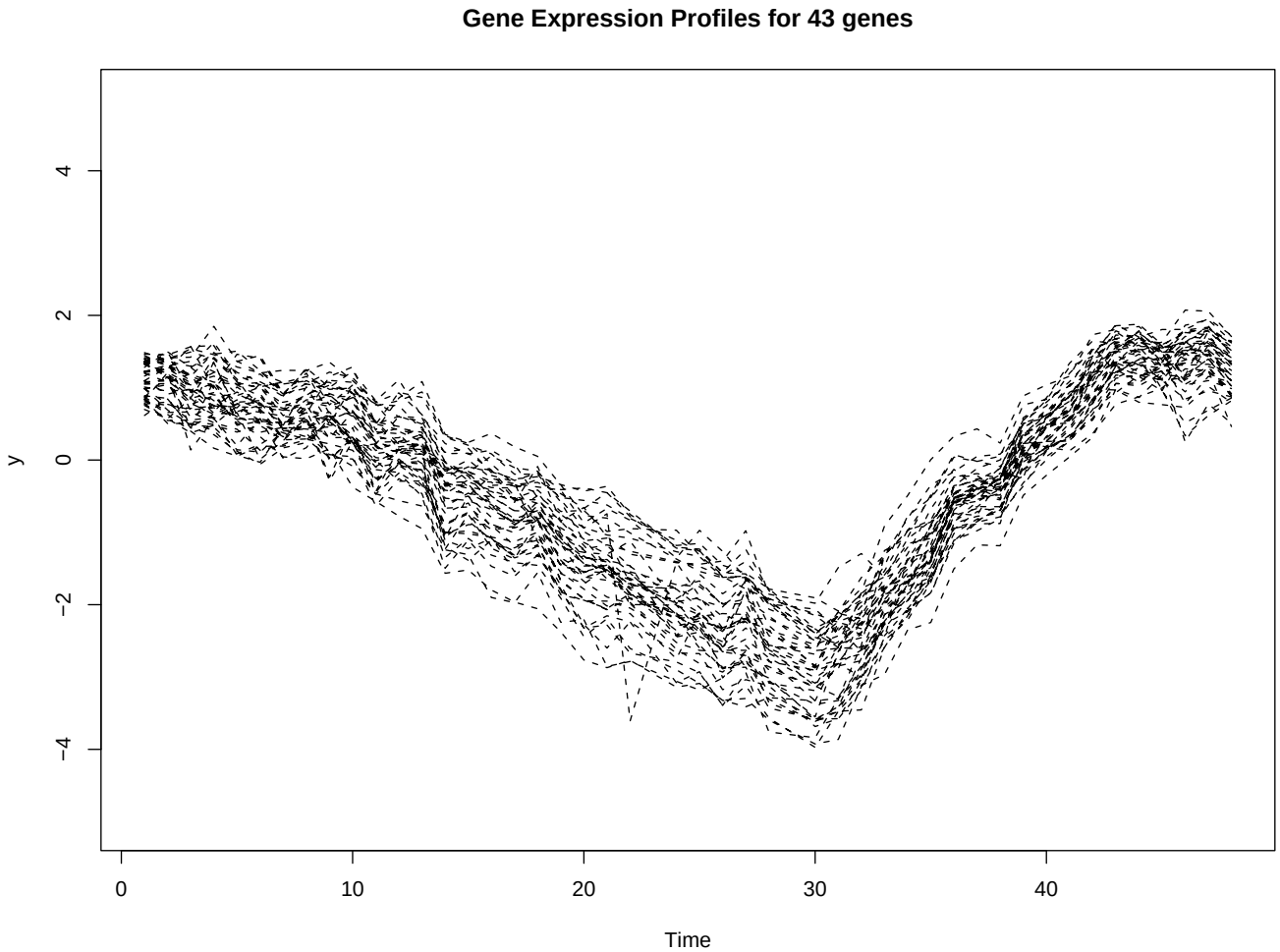
where k is an even number, $k = 2p$ say.

$\lambda_j, j = 1, 2, \ldots, p$ are constants

$$\lambda_1 < \lambda_2 < \cdots < \lambda_p$$

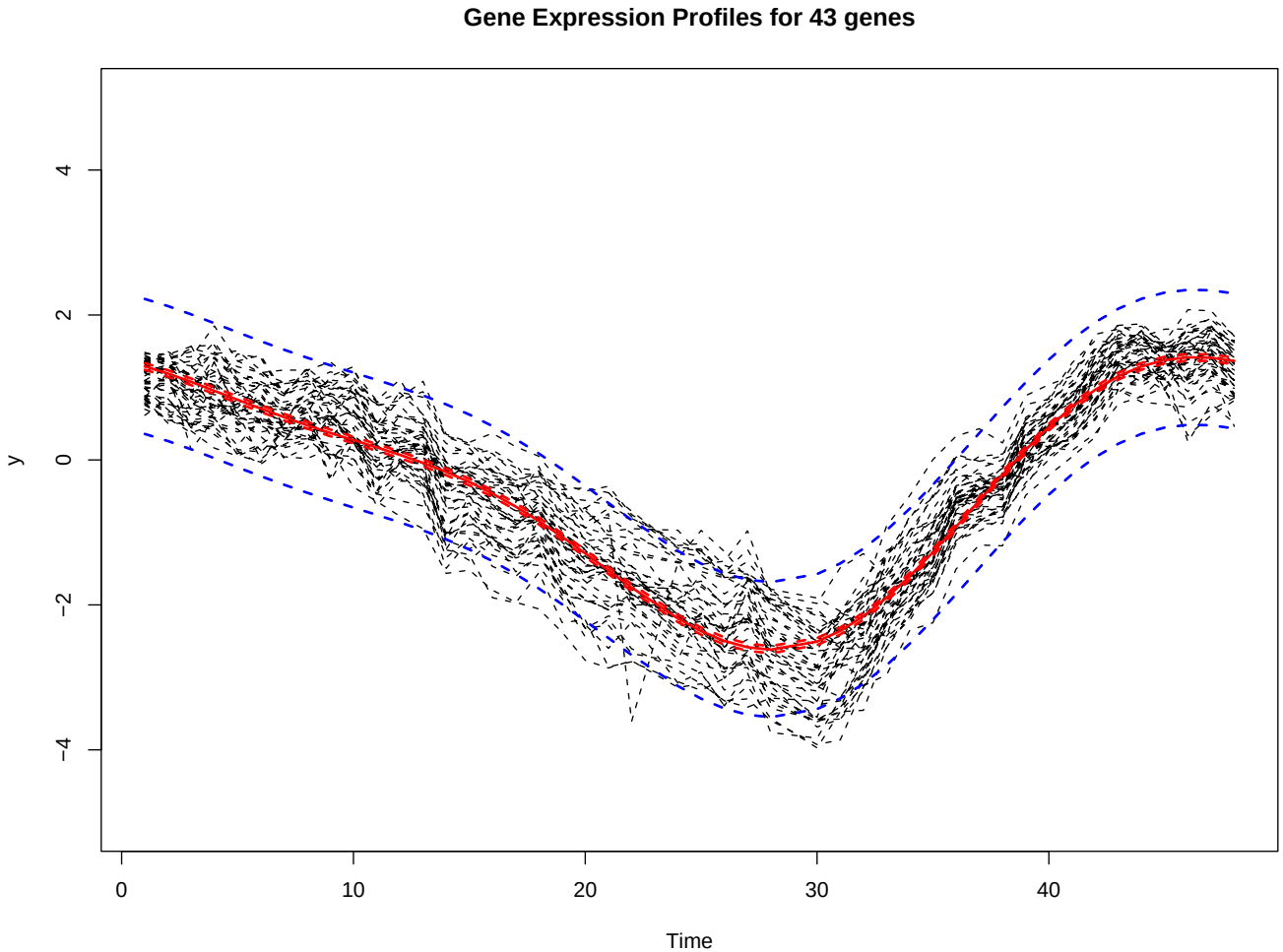
For fixed x, $\cos(\lambda_j x)$ and $\sin(\lambda_j x)$ are also fixed, known values.

Gene expression data



The Linear Model

Harmonic Regression Fit with $p = 2$.



The Linear Model

Calculations are straightforward because the system of equations based on the derivatives

$$\frac{\partial}{\partial \beta_j}\{SSE(\beta)\} = 0 \qquad j = 0, 1, \ldots, k$$

can always be solved routinely, so we can always find $\widehat{\beta}$.

In the general model, simple formulae for

- $\widehat{\beta}$
- e.s.e.$(\widehat{\beta})$
- $\widehat{\sigma}^2$

can be found using a matrix formulation.

Factor predictors

The one-way layout ANOVA can be formulated in the form of the general model.

Recall

- k independent groups
- means $\mu_1, \ldots, \mu_k$
- y_{ij} is the jth observation in the ith group

Factor predictors

Let

$$\begin{aligned} \beta_0 &= \mu_1 \\ \beta_i &= \mu_{i+1} - \mu_1 \qquad i = 1, \ldots, k-1. \end{aligned}$$

Define new predictor $x_{ij}(t)$ where

$$x_{ij}(t) = \begin{cases} 1 & \text{if } t = i \\ 0 & \text{if } t \neq i \end{cases}$$

That is, $x_{ij}(t)$ is one if t matches i (the group label) and is zero otherwise.

Factor predictors

Using the linear regression formulation, we may write

$$Y_{ij} = \beta_0 + \sum_{t=1}^{k-1} \beta_t \mathbf{x}_{ij}(t+1) + \epsilon_{ij}.$$

For any i, j, $x_{ij}(t)$ is non-zero for only one value of t, when $t + 1 = i$.

Therefore, only one term contributes to the summation; the other terms are all zero.

Factor predictors

That is

$$\mathbb{E}[Y_{ij}|x] = \begin{cases} \beta_0 & i = 1 \\ \beta_0 + \beta_1 & i = 2 \\ \beta_0 + \beta_2 & i = 3 \\ \vdots & \\ \beta_0 + \beta_{k-1} & i = k \end{cases}$$

The parameters $\beta_0, \beta_1, \ldots, \beta_{k-1}$ can be estimated using least-squares.

Factor predictors

We term these $x_{ij}(t)$ quantities dummy variables, and the corresponding original variables factor predictors or factors.

This clarifies the link between

ANOVA and Linear Modelling

- they are essentially the same model formulation.

This link extends to all ANOVA models.

Factor predictor example

Milk protein data example:

```
> summary(aov(protein~Diet,data=Milk))

+               Df Sum Sq Mean Sq F value Pr(>F)
+ Diet           2  10.61   5.303   51.85 <2e-16 ***
+ Residuals   1334 136.43   0.102
+ ---
+ Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

> str(Milk)

+ 'data.frame': 1337 obs. of  4 variables:
+  $ protein: num  3.63 3.57 3.47 3.65 3.89 3.73 3.77 3.9 3.78 3.82 ...
+  $ Time   : num  1 2 3 4 5 6 7 8 9 10 ...
+  $ Cow    : Factor w/ 55 levels "B01","B02","B03",..: 1 1 1 1 1 1 1 1 1 1 ...
+  $ Diet   : Factor w/ 3 levels "barley","barley+lupins",..: 1 1 1 1 1 1 1 1 1 1 ...
```

Factor predictor example

```
> summary(lm(protein~Diet,data=Milk))

+
+ Call:
+ lm(formula = protein ~ Diet, data = Milk)
+
+ Residuals:
+      Min       1Q   Median       3Q      Max
+ -0.89193 -0.21969 -0.01238  0.19807  1.16031
+
+ Coefficients:
+                   Estimate Std. Error t value   Pr(>|t|)
+ (Intercept)        3.53193    0.01551 227.681    < 2e-16 ***
+ Dietbarley+lupins -0.10223    0.02153  -4.749 0.00000227 ***
+ Dietlupins        -0.21955    0.02160 -10.166    < 2e-16 ***
+ ---
+ Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
+
+ Residual standard error: 0.3198 on 1334 degrees of freedom
+ Multiple R-squared:  0.07213, Adjusted R-squared:  0.07074
+ F-statistic: 51.85 on 2 and 1334 DF,  p-value: < 2.2e-16
```

Factor predictor example

```
> y<-Milk$protein                      #Response
> x1<-Milk$Diet == 'barley+lupins'     #New predictor for group 2
> x2<-Milk$Diet == 'lupins'            #New predictor for group 3
```

The two variables `x1` and `x2` record the groups corresponding to the factor levels for `Diet`

`barley+lupins` `lupins`

respectively. We then use the R formula

`y ~ x1 + x2`

to represent the summation in the model formula.

Factor predictor example

```
> summary(lm(y~x1+x2))

+
+ Call:
+ lm(formula = y ~ x1 + x2)
+
+ Residuals:
+      Min       1Q   Median       3Q      Max
+ -0.89193 -0.21969 -0.01238  0.19807  1.16031
+
+ Coefficients:
+             Estimate Std. Error t value   Pr(>|t|)
+ (Intercept)  3.53193    0.01551 227.681    < 2e-16 ***
+ x1TRUE      -0.10223    0.02153  -4.749 0.00000227 ***
+ x2TRUE      -0.21955    0.02160 -10.166    < 2e-16 ***
+ ---
+ Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
+
+ Residual standard error: 0.3198 on 1334 degrees of freedom
+ Multiple R-squared:  0.07213, Adjusted R-squared:  0.07074
+ F-statistic: 51.85 on 2 and 1334 DF,  p-value: < 2.2e-16
```

Factor predictor example

That is, the results of the two analyses are identical.

Note that the baseline group – in this case, the `Diet` value `barley` – is not used to define a new predictor; in the `R` formula

```
lm(y~x1+x2)
```

the intercept β_0 is assumed to be present, and corresponds to the baseline group. The parameters

$$\beta_1, \beta_2$$

represents contrasts or differences from baseline.

Multiple Linear Regression

Multiple linear regression models model the variation in response as a function of more than one independent variable.

Suppose we have variables

$$X_1, X_2, \ldots, X_k$$

recording different features of the experimental units. We wish to model response Y as a function of $X_1, X_2, \ldots, X_k$.

Multiple Linear Regression

Consider the model for data point i

$$Y_i = \beta_0 + \beta_1 \mathbf{x}_{i1} + \beta_2 \mathbf{x}_{i2} + \cdots + \beta_k \mathbf{x}_{ik} + \epsilon_i$$

where x_{ij} is the measured value of covariate j on experimental unit i.

Multiple Linear Regression

That is

$$Y_i = \beta_0 + \sum_{j=1}^{k} \beta_j x_{ij} + \epsilon_i$$

The first two terms on the right hand side

$$\beta_0 + \sum_{j=1}^{k} \beta_j x_{ij}$$

are the deterministic components representing the expected value of Y for those x values, and ϵ_i is the random component.

Multiple Linear Regression

Example: $k = 2$

$$Y_i = \beta_0 + \beta_1 \mathbf{x}_{i1} + \beta_2 \mathbf{x}_{i2} + \epsilon_i$$

which is a three parameter model with two predictors.

- β_0 is the intercept
- β_1 represents the contribution of predictor X_1 when X_2 is included in the model
- β_2 represents the contribution of predictor X_2 when X_1 is included in the model

Multiple Linear Regression

We can also include interaction terms

$$Y_i = \beta_0 + \beta_1 x_{i1} + \beta_2 x_{i2} + \beta_{12}(x_{i1} \,.\, x_{i2}) + \epsilon_i$$

where

- The first two terms in x_{i1} and x_{i2} are main effects
- The third term in $(x_{i1} \,.\, x_{i2})$ is an interaction
  - β_{12} represents the amount that X_2 modifies the effect of X_1 when both predictors are included in the model (compared to the case when $X_2 = 0$.

This is a four parameter model.

Multiple Linear Regression

For two continuous predictors X_1 and X_2, the interaction term is obtained by multiplying the two numerical values together

- That is, we compute the product $x_{i1}.x_{i2}$, and include this as a new predictor
- More generally, interactions require more complicated calculations. We will write

$$X_1 : X_2$$

to represent the interaction between X_1 and X_2. This matches R notation.

Multiple Linear Regression Examples

- Multiple regression: Viscosity Example
- Factor Regression
- Interaction
- Residuals

Subgroup analysis

Subgroup analysis is a form of interaction modelling.

Suppose we have

- one factor predictor
- one continuous predictor (or covariate)

where the factor predictor interacts with the covariate so as to modify the slope across the subgroups.

Subgroup analysis

We describe the models using the notation introduced for ANOVA; consider case of

- a single covariate X_1
- a single binary factor predictor X_2

Subgroup analysis

Possible models:

MODEL 0	One horizontal straight line	1
MODEL 1	Two parallel horizontal straight lines	X_2
MODEL 2	One straight line, non-zero slope	X_1
MODEL 3	Two parallel straight lines, with non-zero slope	$X_1 + X_2$
MODEL 4	Two non-parallel straight lines	$X_1 + X_2 + X_1 : X_2$

Subgroup analysis

Note: Always be on the lookout for lurking subgroups (subgroups determined by the levels of an unnoticed factor predictor)

Inferences can change radically when the lurking factor is included in the model

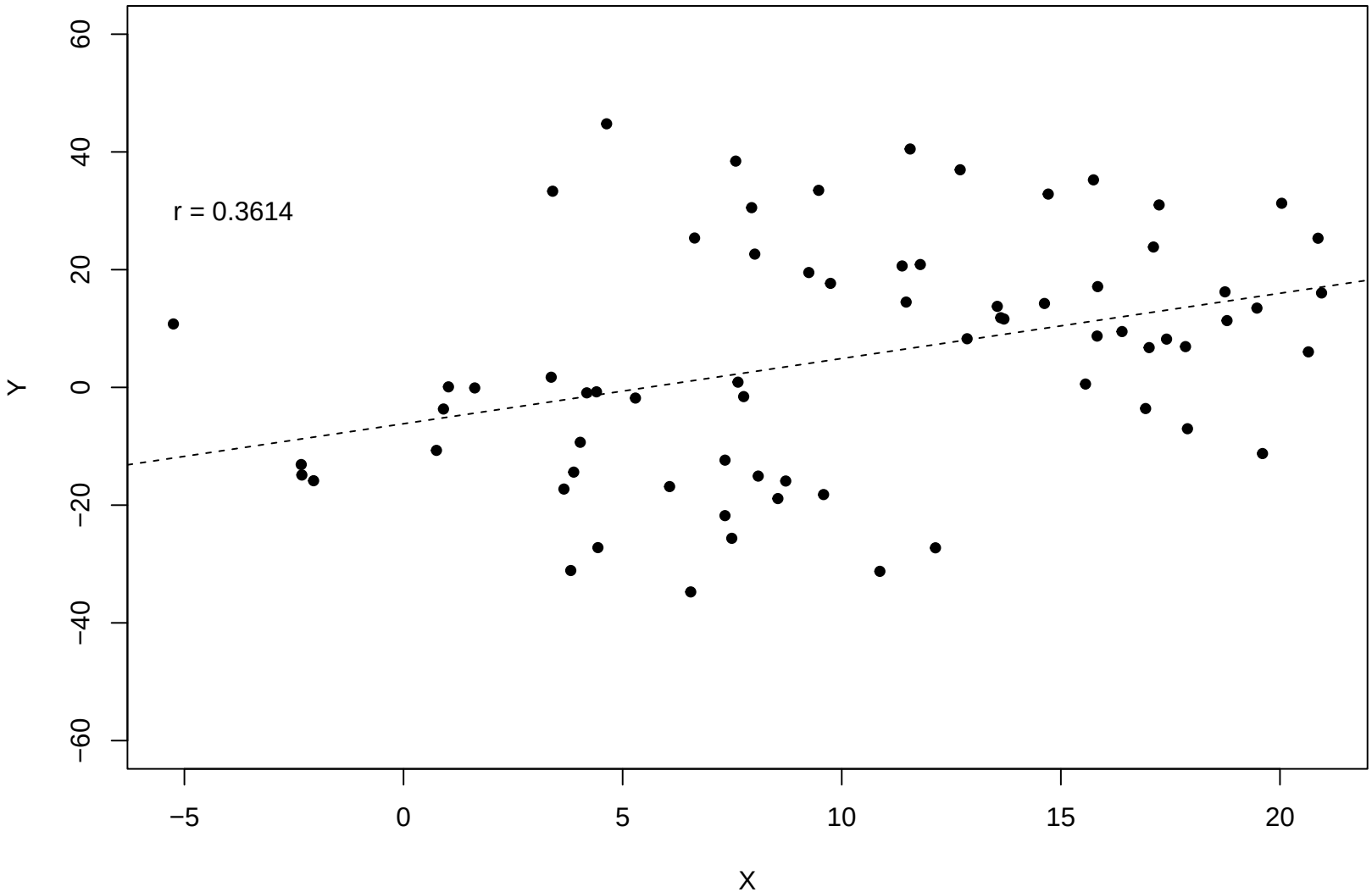
- positive association can be converted into negative association with the continuous covariate.

Subgroup analysis

For example, for factor predictor X_2 taking two levels and continuous covariate X_1.

When the pooled data are examined, a positive association between Y and X_1 is revealed.

Subgroup analysis



Subgroup analysis

If we fit the model

$$\mathbb{E}[Y|\mathbf{x}] = \beta_0 + \beta_1 \mathbf{x}_1$$

the estimate of the slope is positive: $\widehat{\beta}_1 = 1.1081$.

```
> summary(lm(y~x1))


+
+ Call:
+ lm(formula = y ~ x1)
+
+ Residuals:
+     Min      1Q  Median      3Q     Max
+ -37.117 -12.062  -1.197  12.550  45.813
+
+ Coefficients:
+             Estimate Std. Error t value Pr(>|t|)
+ (Intercept)  -6.1727     4.1555  -1.485  0.14206
+ x1            1.1081     0.3467   3.196  0.00211 **
+ ---
+ Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
+
+ Residual standard error: 18.77 on 68 degrees of freedom
+ Multiple R-squared:  0.1306, Adjusted R-squared:  0.1178
+ F-statistic: 10.21 on 1 and 68 DF,  p-value: 0.002114
```

Subgroup analysis

When the pooled data are separated into subgroups,

- $X_2 = 1$ (•) and
- $X_2 = 2$ (+).

a negative association between Y and X_1 in each subgroup is revealed.

Subgroup analysis

Here we fit the model

$$\mathbb{E}[Y|x] = \beta_0 + \beta_1 x_1 + \beta_2 x_{22} + \beta_{12} x_1 . x_{22}$$

where

$$x_{22} = \begin{cases} 1 & \text{for data in Group 2 } (X_2 = 2) \\ 0 & \text{for data in Group 1 } (X_2 = 1) \end{cases}$$

Subgroup analysis

That is, for

- data in Group 1, the model is

$$\mathbb{E}[Y|\mathbf{x}] = \beta_0 + \beta_1 \mathbf{x}_1$$

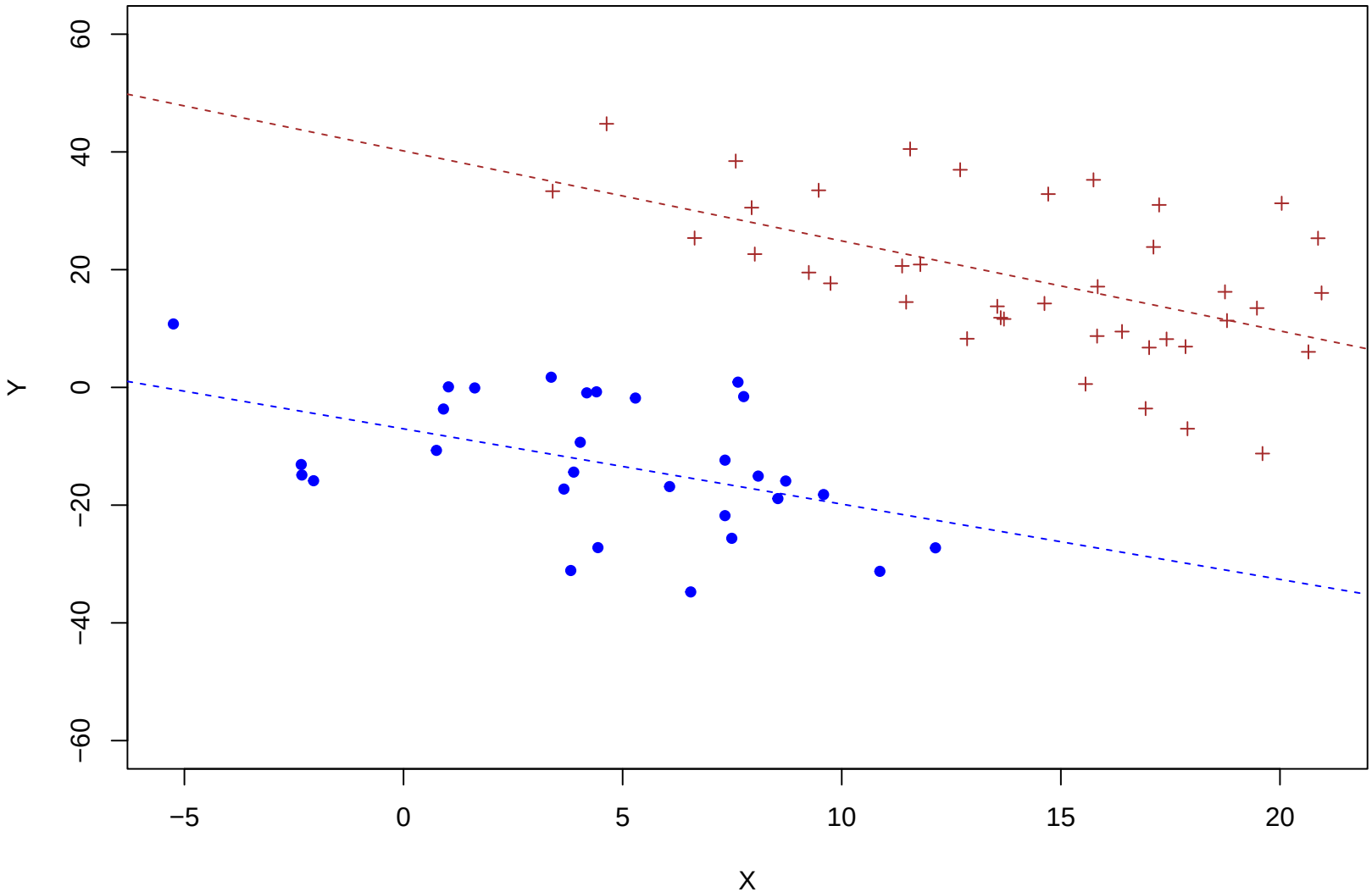
- data in Group 2, the model is

$$\mathbb{E}[Y|\mathbf{x}] = \beta_0 + \beta_1 \mathbf{x}_1 + \beta_2 + \beta_{12} \mathbf{x}_1$$

  more explicitly, a straight line with

$$\text{Intercept: } \beta_0 + \beta_2 \qquad \text{Slope: } \beta_1 + \beta_{12}$$

Subgroup analysis



Subgroup analysis

```
+
+ Call:
+ lm(formula = y ~ x1 + x2 + x1:x2)
+
+ Residuals:
+     Min      1Q  Median      3Q     Max
+ -21.443  -7.361  -1.844   8.879  21.742
+
+ Coefficients:
+             Estimate Std. Error t value Pr(>|t|)
+ (Intercept)  -7.0401     3.0019  -2.345   0.0220 *
+ x1           -1.2787     0.4876  -2.622   0.0108 *
+ x22          47.2124     6.4117   7.363 3.64e-10 ***
+ x1:x22       -0.2502     0.6179  -0.405   0.6868
+ ---
+ Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
+
+ Residual standard error: 10.97 on 66 degrees of freedom
+ Multiple R-squared:  0.7118, Adjusted R-squared:  0.6988
+ F-statistic: 54.35 on 3 and 66 DF,  p-value: < 2.2e-16
```

Subgroup analysis

If we increase X_1, this

- decreases response in subgroup 1
- decreases response in subgroup 2

but appears to increase response overall.

This is known as Simpson's Paradox in Regression. It illustrates that pooling data over subgroups must be carried out with care !

- you must fit the factor predictor in the model if you suspect subgroup differences exist.

Subgroup analysis

In the example, the problem arises due to dependence between X_1 and X_2; all the group with $X_2 = 0$ have low values of X_1, whereas all the group with $X_2 = 1$ have high values of X_1

Dependence between covariates and factor predictors makes model fitting and results interpretation complicated.

Recap

Recap: we can build general models

$$Y_i = \beta_0 + \sum_{j=1}^{k} \beta_j \mathbf{x}_{ij} + \epsilon_i$$

to explain the variation of y in terms of covariates and factor predictors $x_1, \ldots, x_k$.

- Simple Linear Regression
- Polynomial Regression
- Multiple Regression
- Factor Predictor Regression
- Interaction Models

Recap

We can fit each of these models easily using least-squares to obtain

- estimates $\widehat{\beta} = (\widehat{\beta}_1, \widehat{\beta}_2, \ldots, \widehat{\beta}_k)^\top$
- standard errors
- goodness of fit measures R^2 and Adjusted R^2
- residuals for model checking
- predictions

Interpreting $\widehat{\beta}_j$

$\widehat{\beta}_j$ can be interpreted as the amount of increase in response y when x_j increases by one unit when the other predictors

$$x_1, x_2, \ldots, x_{j-1}, x_{j+1}, \ldots, x_k$$

are held fixed.

We can test the hypothesis

$$\begin{aligned} H_0 &: \beta_j = 0 \\ H_0 &: \beta_j \neq 0 \end{aligned}$$

using the usual hypothesis testing approach.

Interpreting $\widehat{\beta}_j$

Test statistic:

$$T_{obs,j} = \frac{\widehat{\beta}_j}{\text{e.s.e.}(\widehat{\beta}_j)} = \frac{\text{ESTIMATE}}{\text{STANDARD ERROR}}$$

If H_0 is true,

$$T_{obs,j} \sim t_{n-k-1}$$

as we are estimating $k+1$ parameters overall.

Note: In multiple regression, when testing each of

$$\widehat{\beta}_0, \widehat{\beta}_1, \ldots, \widehat{\beta}_k$$

we should strictly use a multiple testing correction (as in post-hoc tests in ANOVA)

Model Checking

We can fit models with different numbers of predictors, and

- assess whether any individual covariate is influential in the model (look at t-statistics).
- assess whether there is any explanatory power in the variables combined (look at ANOVA statistics)

ANOVA table for multiple regression

SOURCE	DF	SS	MS	F
REGRESSION	k	SSR	MSR	$F = \dfrac{MSR}{MSE}$
ERROR	$n-k-1$	SSE	MSE	
TOTAL	$n-1$	SS		

where

$$MSR = \frac{SSR}{k} \qquad MSE = \frac{SSE}{n-k-1}$$

the F-statistic is

$$F = \frac{MSR}{MSE}$$

ANOVA table for multiple regression

If H_0 is true

$$F \sim F_{k,n-k-1}.$$

Here

$$\begin{aligned} H_0 &: \beta_1 = \beta_2 = \cdots = \beta_k = 0 \\ H_a &: \text{At least one } \beta_j \neq 0 \end{aligned}$$

- The model for H_0 has one parameter β_0.
- The model for H_a has $k + 1$ parameters

$$\beta_0, \beta_1, \beta_2, \ldots, \beta_k$$

ANOVA table for multiple regression

- Therefore the number of extra parameters for model H_a is

$$(k+1) - 1 = k$$

  To obtain model H_0 from model H_a we constrain k parameters to be zero.

ANOVA table for multiple regression

Because we can constrain model H_a by setting some parameters to zero to obtain model H_0, we say that

Model H_0 is nested inside Model H_a

The number, k, of constraints $\beta_1 = \beta_2 = \cdots = \beta_k = 0$ explains why the ANOVA table Regression degrees of freedom is k

- the multiple regression brings in k extra parameters.

ANOVA table for multiple regression

In addition, we can use the R^2 or Adjusted R^2 statistic to check overall model adequacy

$$R^2 = 1 - \frac{SSE}{SS_{yy}} = \frac{SS_{yy} - SSE}{SS_{yy}} = \frac{SSR}{SS}$$

which is equal to

$$\frac{\text{VARIATION EXPLAINED BY THE REGRESSION}}{\text{TOTAL VARIATION}}$$

Also

$$R^2_{adj} = 1 - \frac{SSE/(n-k-1)}{SS/(n-1)}$$

$R^2 > 0.7$ implies that the model is a good fit, that is, most of the variation observed is explained by the regression model.

ANOVA table for multiple regression

We can now fit completely general models in the form of the Linear Model; if y is the response, and $x_1, \ldots, x_k$ are the covariates or factor predictors, we can include combinations of

- Polynomial Main Effects : $x_j, x_j^2, x_j^3, \ldots$
- Two-way Interactions: $x_{j_1}.x_{j_2}$
- Three-way Interactions: $x_{j_1}.x_{j_2}.x_{j_3}$

etc.

Dummy Variables

Recall that we can fit a factor predictor using dummy variables.

For example, if factor predictor X has L levels, we pick one as the baseline level, and then create $L-1$ new binary predictors $X_1, \ldots, X_{L-1}$, where, for $l = 1, \ldots, L-1$

$$X_l = \begin{cases} 1 & \text{whenever } X = l \\ 0 & \text{otherwise} \end{cases}$$

We can then include $X_1, \ldots, X_{L-1}$ in the regression model, along with an intercept which represents the mean level for the baseline group.

Dummy Variables

Example: $L = 4$

We pick $X = 1$ as the baseline group, and define X_1, X_2 and X_3 as follows:

X	X_1	X_2	X_3
3	0	1	0
1	0	0	0
3	0	1	0
4	0	0	1
2	1	0	0
2	1	0	0

See textbook Section 12.7.

Stepwise Model Selection

We seek a method that allows us to compare nested models.

Suppose we want to compare

$$\begin{aligned} \text{MODEL 1} \quad &: \quad \mathbb{E}[Y|x] = \beta_0 + \beta_1 x + \beta_2 x^2 \\ \text{MODEL 2} \quad &: \quad \mathbb{E}[Y|x] = \beta_0 + \beta_1 x + \beta_2 x^2 + \beta_3 x^3 \end{aligned}$$

Model 1 is nested inside Model 2 as if we set $\beta_3 = 0$ in Model 2, we get Model 1.

Stepwise Model Selection

If

$$\begin{aligned} \text{MODEL 1} \quad &: \quad \mathbb{E}[Y|x] = \beta_0 + \beta_1 x_1 + \beta_2 x_2 \\ \text{MODEL 2} \quad &: \quad \mathbb{E}[Y|x] = \beta_0 + \beta_1 x_1 + \beta_2 x_2 + \beta_{12}(x_1 . x_2) \end{aligned}$$

we can set $\beta_{12} = 0$ in Model 2 to obtain Model 1, so again the models are nested.

We can set up a hypothesis test to assess whether the simplification of Model 2 to Model 1 (by setting one or more parameters equal to zero) is justified by the data.

ANOVA tests for Comparing Nested Models

Terminology:

- Complete Model

$$\mathbb{E}[Y|x] = \beta_0 + \beta_1 x_1 + \cdots + \beta_k x_k$$

- Reduced Model

$$\mathbb{E}[Y|x] = \beta_0 + \beta_1 x_1 + \cdots + \beta_g x_g$$

where $g < k$. The reduced model is obtained from the complete model by setting

$$\beta_{g+1} = \beta_{g+2} = \cdots = \beta_k = 0$$

ANOVA tests for Comparing Nested Models

The reduced model is nested inside the complete model.

We wish to test the hypothesis

$$\begin{aligned} H_0 &: \beta_{g+1} = \beta_{g+2} = \cdots = \beta_k = 0 \\ H_a &: \text{At least one of these } \beta_j \neq 0 \end{aligned}$$

We can test this hypothesis by fitting both models, and combining the results; we focus on the sums of squares quantities.

Model comparison method

1. Fit the complete model and obtain the sum of squared errors, SSE_C, available from the ANOVA table.
2. Fit the reduced model and obtain the sum of squared errors, SSE_R, available from the ANOVA table.
3. Form the test statistic

$$F = \frac{(SSE_R - SSE_C)/(k-g)}{SSE_C/(n-k-1)}$$

Model comparison method

If H_0 is true, then $F \sim F_{k-g,n-k-1}$.

- $k - g$ is the number of parameters we set equal to zero when moving from complete to reduced model.
- Using this F statistic, we can assess whether there is evidence to support the reduced model over the complete model

Model comparison method

Complete Model ANOVA table:

SOURCE	DF	SS	MS	F
COMPLETE MODEL	k	SSR_C	MSR_C	F_C
ERROR$_C$	$n-k-1$	SSE_C	MSE_C	
TOTAL	$n-1$	SS		

Model comparison method

Reduced Model ANOVA table:

SOURCE	DF	SS	MS	F
REDUCED MODEL	g	SSR_R	MSR_R	F_R
ERROR$_R$	$n-g-1$	SSE_R	MSE_R	
TOTAL	$n-1$	SS		

Model comparison method

The result holds for comparing any two nested models where the standard model assumptions hold:

- ϵ uncorrelated
- ϵ independent of $x_1, \ldots, x_k$
- ϵ has constant variance
- $\epsilon \sim Normal(0, \sigma^2)$

Model comparison method

We cannot use this approach to compare non-nested models; for example

$$\begin{aligned} \text{MODEL 1} &: Y = \beta_0 + \beta_1 \mathbf{x}_1 + \epsilon \\ \text{MODEL 2} &: Y = \beta_0 + \beta_2 \mathbf{x}_2 + \epsilon \end{aligned}$$

cannot be compared using the F-test.

The statistic

$$F = \frac{(SSE_R - SSE_C)/(k-g)}{SSE_C/(n-k-1)} = \frac{①/②}{③/④}$$

① - $SSE_R - SSE_C$: this is the improvement in fit when the reduced model is extended to the complete model

② - $k - g$: this is the number of extra parameters needed to extend the reduced model to the complete model

③ - SSE_C

④ - $n - k - 1$

The statistic

In the numerator

$$① = SSE_R - SSE_C$$

is always positive. This is because we constrain the Complete model to obtain the Reduced model, so the Reduced model is less flexible to fit to the data.

The statistic

In the denominator,

$$\frac{③}{④} = \frac{SSE_C}{n-k-1}$$

is the best guess we have at the true value of σ^2, that is, the estimate of σ^2 constructed using as much information as possible, once the effects of

$$x_1, \ldots, x_k$$

have been accounted for.

Example: Hooker's data

Example: Hooker's Data

We consider the two models:

$$\begin{aligned} \text{MODEL 1} &: Y = \beta_0 + \beta_1 x + \epsilon \\ \text{MODEL 2} &: Y = \beta_0 + \beta_1 x + \beta_2 x^2 + \epsilon \end{aligned}$$

Here

- MODEL 1: Reduced Model
- MODEL 2: Complete Model

$k = 2$, $g = 1$.

IS THE QUADRATIC TERM NEEDED ?

Example: Hooker's data

Example: Hooker's Data: straight-line

```
> supernova(lm(temp~pressure, data=Hooker), type=2)

+  Analysis of Variance Table (Type II SS)
+  Model: temp ~ pressure
+
+                              SS df       MS        F   PRE     p
+  ----- --------------- | -------- -- -------- -------- ----- -----
+  Model (error reduced) | 2272.474  1 2272.474 3497.902 .9918 .0000
+  Error (from model)    |   18.840 29    0.650
+  ----- --------------- | -------- -- -------- -------- ----- -----
+  Total (empty model)   | 2291.315 30   76.377
```

Example: Hooker's data

Example: Hooker's Data: quadratic

```
> supernova(lm(temp~pressure+I(pressure^2), data=Hooker), type=2)

+  Analysis of Variance Table (Type II SS)
+  Model: temp ~ pressure + I(pressure^2)
+
+                                      SS df       MS        F   PRE     p
+  ------------ --------------- | -------- -- -------- -------- ----- -----
+          Model (error reduced) | 2286.933  2 1143.467 7306.975 .9981 .0000
+       pressure                 |   68.260  1   68.260  436.197 .9397 .0000
+  I(pressure^2)                 |   14.459  1   14.459   92.394 .7674 .0000
+          Error (from model)    |    4.382 28    0.156
+  ------------ --------------- | -------- -- -------- -------- ----- -----
+          Total (empty model)   | 2291.315 30   76.377
```

Example: Hooker's data

Example: Hooker's Data

COMPLETE MODEL	SSR_C	2286.933
	SSE_C	4.382
REDUCED MODEL	SSR_R	2272.474
	SSE_R	18.840

with $n = 31, k = 2, g = 1$, so $k - g = 1$ and $n - k - 1 = 28$.

Example: Hooker's data

Example: Hooker's Data

Hence

$$F_{obs} = \frac{(SSE_R - SSE_C)/(k-g)}{SSE_C/(n-k-1)}$$

$$= \frac{(18.840 - 4.382)/1}{4.382/28}$$

$$= 92.383$$

Example: Hooker's data

Example: Hooker's Data

We compare F with the

$$F_{k-g,n-k-1} \equiv F_{1,28}$$

distribution. The upper $\alpha = 0.05$ point is

$$F_{0.05,1,28} = 4.20$$

Thus

$$92.383 = F_{obs} > F_{0.05,1,28} = 4.20$$

and H_0 is rejected in favour of H_a.

The quadratic model fits better than the straight-line model.

Example: Hooker's data

From the original ANOVA tables, we already know that Model 1 and Model 2 both fit better than the null model

$$\begin{aligned} \text{MODEL 0} \qquad \mathbb{E}[Y|\mathbf{x}] &= \beta_0 \\ Y &= \beta_0 + \epsilon \end{aligned}$$

where there is no dependence on x.

We have now also confirmed that Model 2 fits better than Model 1.

Diabetes data

Example: Diabetes Data

Factor Predictor: group (X_2)

Continuous Covariate: loggluf (X_1)

Response: logglut (Y)

We have five models to confirm:

$$\begin{aligned}
\text{MODEL 0} &: 1 \\
\text{MODEL 1} &: X_2 \\
\text{MODEL 2} &: X_1 \\
\text{MODEL 3} &: X_1 + X_2 \\
\text{MODEL 4} &: X_1 + X_2 + X_1 : X_2
\end{aligned}$$

Diabetes data

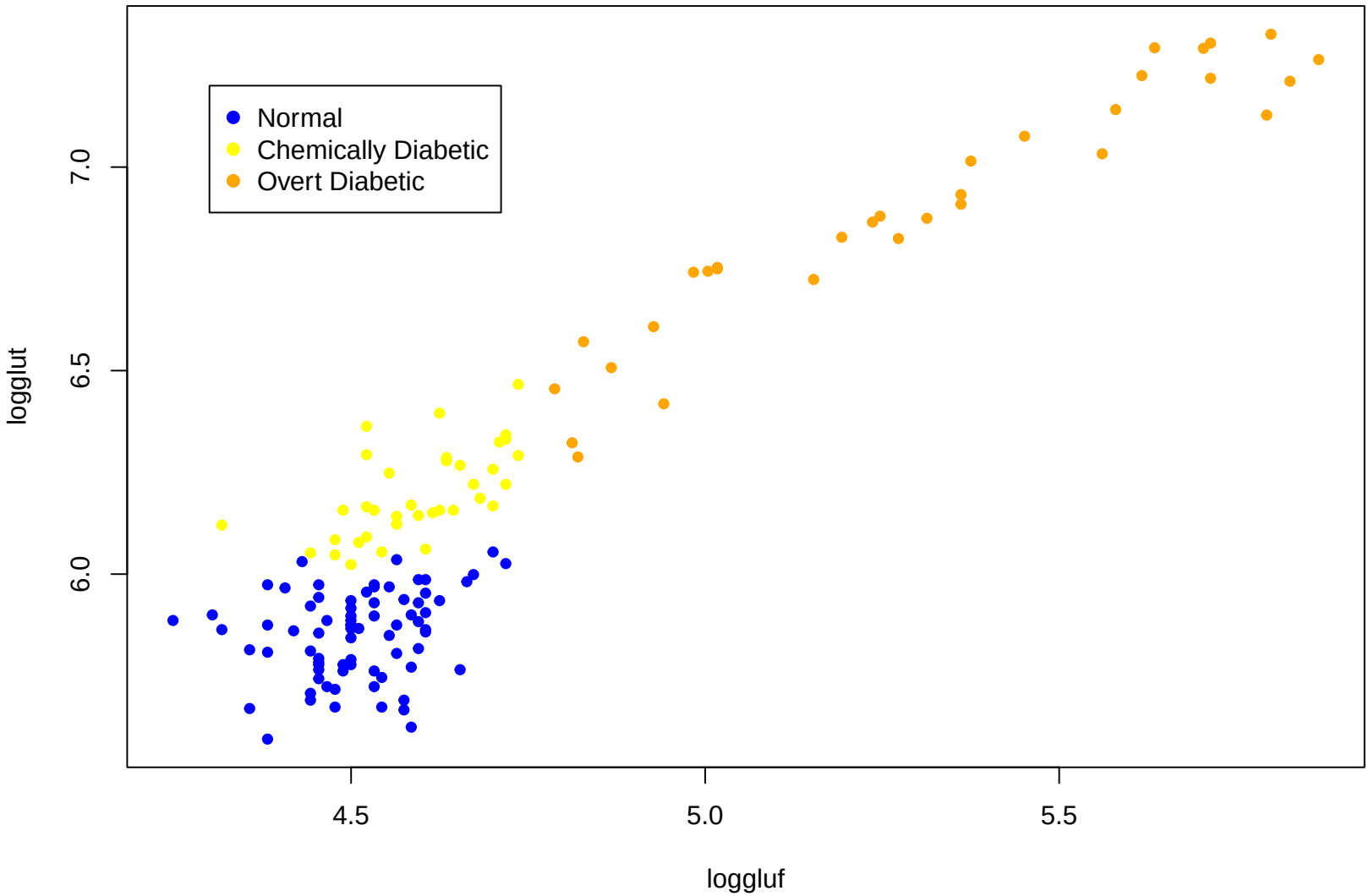
```
> str(Diabetes)

+ 'data.frame': 144 obs. of  15 variables:
+  $ id          : num  1 2 3 4 5 6 7 8 9 10 ...
+  $ relwt       : num  0.81 0.95 0.94 1.04 1 0.76 0.91 1.1 0.99 0.78 ...
+  $ glufast     : num  80 97 105 90 90 86 100 85 97 97 ...
+  $ glutest     : num  356 289 319 356 323 381 350 301 379 296 ...
+  $ instest     : num  124 117 143 199 240 157 221 186 142 131 ...
+  $ sspg        : num  55 76 105 108 143 165 119 105 98 94 ...
+  $ group       : Factor w/ 3 levels "Overt Diabetic",..: 3 3 3 3 3 3 3 3 3 3 ...
+  $ loggluf     : num  4.38 4.57 4.65 4.5 4.5 ...
+  $ logglut     : num  5.87 5.67 5.77 5.87 5.78 ...
+  $ normal      : num  1 1 1 1 1 1 1 1 1 1 ...
+  $ overt       : num  0 0 0 0 0 0 0 0 0 0 ...
+  $ chemical    : num  0 0 0 0 0 0 0 0 0 0 ...
+  $ normalgluf  : num  4.38 4.57 4.65 4.5 4.5 ...
+  $ overtgluf   : num  0 0 0 0 0 0 0 0 0 0 ...
+  $ chemicalgluf: num  0 0 0 0 0 0 0 0 0 0 ...

> table(Diabetes$group)

+
+      Overt Diabetic Chemically Diabetic              Normal
+                  32                   36                  76
```

Diabetes data



Diabetes data

Example: Diabetes Data

MODEL 4 is the most complex model with 6 parameters

$$\beta_{10}, \beta_{11}, \beta_{20}, \beta_{21}, \beta_{30}, \beta_{31}$$

MODEL 4:

$$\mathbb{E}[Y|x] = \begin{cases} \beta_{10} + \beta_{11} x_1 & \text{GROUP 1} \\ \beta_{20} + \beta_{21} x_1 & \text{GROUP 2} \\ \beta_{30} + \beta_{31} x_1 & \text{GROUP 3} \end{cases}$$

All of the other models are nested inside Model 4; we can obtain them all by setting parameters equal to zero.

Diabetes data

Example: Diabetes Data

In the default parameterization:

β_{10}, β_{11} Group 1 Intercept and Slope

$\beta_{20} = \beta_{10} + \delta_{20}$ Changes in the Intercepts for
$\beta_{30} = \beta_{10} + \delta_{30}$ groups 2 and 3 are δ_{20} and δ_{30}

$\beta_{21} = \beta_{11} + \delta_{21}$ Changes in the Slopes fpr
$\beta_{31} = \beta_{11} + \delta_{31}$ groups 2 and 3 are δ_{21} and δ_{31}

Thus the six new parameters are

$$\beta_{10}, \beta_{11}, \delta_{20}, \delta_{30}, \delta_{21}, \delta_{31}$$

Diabetes data

MODEL 0 $\beta_{11} = 0$
$\delta_{20} = \delta_{30} = \delta_{21} = \delta_{31} = 0$

MODEL 1 $\beta_{11} = \delta_{21} = \delta_{31} = 0$

MODEL 2 $\delta_{20} = \delta_{30} = \delta_{21} = \delta_{31} = 0$

MODEL 3 $\delta_{21} = \delta_{31} = 0$

Note: $\beta_{11} = 0 \Longrightarrow \delta_{21} = \delta_{31} = 0$, as X_1 is not included in the model.

Counting Parameters

Whenever we remove a

- continuous covariate, we set one parameter to zero.
- factor predictor with L levels, we set $L - 1$ parameters to zero.
- two-way interaction between the variables, we set

$$1 \times (L - 1) = L - 1$$

parameters to zero.

Counting Parameters

Models 0,1,2,3 are nested inside Model 4.

Two approaches to finding the best model are used:

1. Start with Model 0 and try to add terms that improve the model fit (Forward Selection)
2. Start with Model 4 and try to remove terms that improve the model fit (Backward Selection)

Note:

- Models 0,1 and 2 are nested inside Model 3.
- Model 0 is nested inside Models 1 and 2.

Therefore we can begin with Model 4, or Model 3 or Model 1 or 2, and simplify to a nested model.

Example: Diabetes data

```
> fit0<-lm(logglut~1, data=Diabetes)                                  #Model 0
> fit1<-lm(logglut~group, data=Diabetes)                              #Model 1
> fit2<-lm(logglut~loggluf, data=Diabetes)                            #Model 2
> fit3<-lm(logglut~loggluf+group, data=Diabetes)                      #Model 3
> fit4<-lm(logglut~loggluf+group+loggluf:group, data=Diabetes)        #Model 4
```

SSE quantities:

```
> SSE.0<-sum(residuals(fit0)^2)
> SSE.1<-sum(residuals(fit1)^2)
> SSE.2<-sum(residuals(fit2)^2)
> SSE.3<-sum(residuals(fit3)^2)
> SSE.4<-sum(residuals(fit4)^2)
> c(SSE.0,SSE.1,SSE.2,SSE.3,SSE.4)

+ [1] 28.504333  4.159984  3.737941  1.472320  1.317650
```

Example: Diabetes data

Example: Diabetes Data

Here $n = 144$.

Model	Description	SSE	p
0	1	28.504	1
1	X_2	4.160	3
2	X_1	3.738	2
3	$X_1 + X_2$	1.472	4
4	$X_1 + X_2 + X_1 : X_2$	1.318	6

p is the number of non-zero parameters; k or g is always $p-1$ in the following analysis.

Example: Diabetes data

Backward Selection:

Complete Model : Model 4
Reduced Model : Model 3

Here $k = 5, g = 3$ so $k - g = 2$, and

$$n - k - 1 = 144 - 5 - 1 = 138.$$

We have

$$F_{obs} = \frac{(SSE_R - SSE_C)/(k-g)}{SSE_C/(n-k-1)} = \frac{(1.472 - 1.318)/2}{1.318/138} = 8.062$$

Example: Diabetes data

We compare this with the

$$F_{k-g,n-k-1} = F_{2,138}$$

distribution; we have $F_{0.05,2,138} = 3.061$, so we

Reject H_0 at $\alpha = 0.05$.

Example: Diabetes data

That is, Model 4

$$X_1 + X_2 + X_1 : X_2$$

fits significantly better than Model 3

$$X_1 + X_2.$$

We cannot simplify the complete model to the reduced model without the loss of significant explanatory power.

The Interaction is Necessary in the Model

Backward selection stops here; we cannot simplify further from the complete model.

Example: Diabetes data

```
> anova(fit3, fit4, test='F')

+ Analysis of Variance Table
+
+ Model 1: logglut ~ loggluf + group
+ Model 2: logglut ~ loggluf + group + loggluf:group
+   Res.Df    RSS Df Sum of Sq      F    Pr(>F)
+ 1    140 1.4723
+ 2    138 1.3176  2   0.15467 8.0994 0.0004721 ***
+ ---
+ Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

> qf(0.95, 2, 138)

+ [1] 3.061716
```

Example: Diabetes data

Forward Selection: we start with Model 0 and build up.

Model 1 vs Model 0 $\quad F_{obs} = 412.568$

Model 2 vs Model 0 $\quad F_{obs} = 940.846$

It seems that Model 2 is the better improvement, so we try the selection path

$$\text{Model 0} \longrightarrow \text{Model 2} \longrightarrow \text{Model 3} \longrightarrow \text{Model 4}$$

Example: Diabetes data

Model	SSE	$SSE_R - SSE_C$
0	28.504	-
2	3.738	24.766
3	1.472	2.266
4	1.318	0.154

We work down the table, $28.504 - 3.738 = 24.766$ etc.

Example: Diabetes data

Comparison	k	g	SSE_C	$SSE_R - SSE_C$	F_{obs}
2 vs 0	1	0	3.738	24.766	940.82
3 vs 2	3	1	1.472	2.266	107.76
4 vs 3	5	3	1.318	0.154	8.06

Recall that $n = 144$, and

$$F = \frac{(SSE_R - SSE_C)/(k-g)}{SSE_C/(n-k-1)}$$

Under each H_0,

$$F \sim F_{k-g,n-k-1}.$$

Example: Diabetes data

- $F_{0.05,1,142} \doteq 3.92 < 940.82$
  Therefore Model 0 is NOT an adequate simplification of Model 2
- $F_{0.05,2,140} \doteq 3.07 < 107.76$
  Therefore Model 2 is NOT an adequate simplification of Model 3
- $F_{0.05,2,138} \doteq 3.07 < 8.06$
  Therefore Model 3 is NOT an adequate simplification of Model 4

All of the null hypotheses are rejected.

Example: Diabetes data

By both forward and backward selection, we select Model 4

$$X_1 + X_2 + X_1 : X_2$$

as the most appropriate model.

Here the convention is not to correct for multiple testing (we don't know how many tests we are going to do), although a correction could be used.

F-tests for Unbalanced Designs

Example: Potato Damage Data

The damage to potato plants caused by cold temperatures is to be studied.

In this experimental study, three binary factor predictors were used

- we label them A, B and C rather than X_1, X_2, X_3 to recall the link with Factorial Designs in ANOVA.

F-tests for Unbalanced Designs

Each factor takes two levels:

Example: Potato Damage Data

	Factor	Levels
A	Potato Variety	0- Variety 1, 1- Variety 2
B	Acclimatization Routine	0- Room Temp, 1- Cold Room
C	Preparation Treatment	0- -4C, 1- -8C

Thus we have a $2 \times 2 \times 2$ three-way factorial design.

F-tests for Unbalanced Designs

However, the design is unbalanced; we have different numbers of replicates in each of the 8 factor-level combinations.

- we cannot use conventional 3-way ANOVA; the lack of balance means that the stated p-valuesm ay be misleading if we perform a standard ANOVA.

- we are forced to use the F-test model comparison approach.

F-tests for Unbalanced Designs

```
> str(Potato)

+ 'data.frame': 75 obs. of  5 variables:
+  $ potato: Factor w/ 2 levels "Variety 1","Variety 2": 1 1 1 1 1 1 1 1 1 1 ...
+  $ regime: Factor w/ 2 levels "Room Temp","Cold Room": 1 1 1 1 1 1 1 1 1 1 ...
+  $ temp  : Factor w/ 2 levels " -4 C"," -8 C": 1 1 1 1 1 2 2 2 2 2 ...
+  $ photo : num  4.64 5.79 29.35 18.42 2.26 ...
+  $ leak  : num  2.25 4.34 4.25 6.14 2.38 16.3 5.24 3.25 0.88 3.98 ...

> ftable(xtabs(~potato+regime+temp,data=Potato))

+                     temp  -4 C  -8 C
+ potato    regime
+ Variety 1 Room Temp          5     5
+           Cold Room         12    13
+ Variety 2 Room Temp         13    13
+           Cold Room          7     7
```

There are unequal numbers of replicates in the 8 treatments.

F-tests for Unbalanced Designs

```
> supernova(lm(leak~potato*regime*temp,data=Potato),type=2)

+  Analysis of Variance Table (Type II SS)
+  Model: leak ~ potato * regime * temp
+
+                                          SS df       MS      F   PRE     p
+  ----------------- --------------- | --------- -- -------- ------ ----- -----
+               Model (error reduced) |  8842.339  7 1263.191 17.033 .6402 .0000
+              potato                 |  1503.499  1 1503.499 20.273 .2323 .0000
+              regime                 |  1973.358  1 1973.358 26.609 .2843 .0000
+                temp                 |  1225.714  1 1225.714 16.527 .1979 .0001
+       potato:regime                 |  2093.162  1 2093.162 28.224 .2964 .0000
+         potato:temp                 |    35.447  1   35.447  0.478 .0071 .4917
+         regime:temp                 |     6.791  1    6.791  0.092 .0014 .7631
+  potato:regime:temp                 |    89.198  1   89.198  1.203 .0176 .2767
+               Error (from model)    |  4968.876 67   74.162
+  ----------------- --------------- | --------- -- -------- ------ ----- -----
+               Total (empty model)   | 13811.215 74  186.638
```

The ANOVA F-tests are not reliable due to the unbalanced design, so we must use the nested model F-tests.

F-tests for Unbalanced Designs

Here the most complex model can be written

$$A + B + C + A : B + A : C + B : C + A : B : C$$

that is,

- all main effects (terms 1,2 and 3)
- all two-way interactions (terms 4,5 and 6)
- all three-way interactions (term 7)

We may write this model

$$A * B * C$$

which is termed the full factorial model.

Counting the numbers of parameters

Term	Parameters	
A	$(a-1)$	1
B	$(b-1)$	1
C	$(c-1)$	1
$A:B$	$(a-1)(b-1)$	1
$A:C$	$(a-1)(c-1)$	1
$B:C$	$(b-1)(c-1)$	1
$A:B:C$	$(a-1)(b-1)(c-1)$	1
Total		7

where $a = b = c = 2$.

We have 7 parameters in total (excluding the baseline mean) when all terms are considered, so

$$k = 7$$

Counting the numbers of parameters

In the following tables columns are:

- Complete Model
- Reduced Model
- SSE_C
- SSE_R
- k
- g
- F_{obs} (test statistic)
- $F_{0.05,k-g,n-k-1}$

We denote the critical value by F_α and check whether $F > F_\alpha$.

Potato Damage Data: ANOVA F-Tests

We compare four models: M_{R_1}, M_{R_2} and M_{R_3} are nested within the complete model M_C.

$$
\begin{aligned}
M_C &: A + B + C + A:B + A:C + B:C + A:B:C \\
M_{R_1} &: A + B + C + A:B \\
M_{R_2} &: A + B + C \\
M_{R_3} &: A + B + A:B
\end{aligned}
$$

Potato Damage Data: ANOVA F-Tests

C	R	SSE_C	SSE_R	k	g	F_{obs}	F_α
M_C	M_{R_1}	4968.876	5093.746	7	4	0.561	2.76
M_{R_1}	M_{R_2}	5093.746	7183.674	4	3	28.721	3.92
M_{R_1}	M_{R_3}	5093.746	6319.640	4	3	16.846	3.92

Potato Damage Data: ANOVA F-Tests

Note: The quoted F_α values are approximate as the textbook does not tabulate all F distributions. We take $\alpha = 0.05$

Conclusions

Taking the comparisons in order:

1. M_C *vs* M_{R_1} : $F < F_\alpha$. This result is not significant: Model M_{R_1} is an adequate simplification of Model M_C, and we choose M_{R_1} over M_C.

   The model M_{R_1} now becomes the complete model.

2. M_{R_1} *vs* M_{R_2} : $F > F_\alpha$. This result is significant: Model M_{R_2} is not an adequate simplification of Model M_{R_1}

3. M_{R_1} *vs* M_{R_3} : $F > F_\alpha$. This result is significant: Model M_{R_3} is not an adequate simplification of Model M_{R_1}

Conclusions

Thus the final model is

$$A + B + C + A : B$$

i.e. all main effects, plus the interaction between potato variety and acclimatization routine.

We cannot simplify this model further without significant loss in terms of goodness of fit.

Note: $R^2 = 0.631$ and $R^2_{adj} = 0.610$, so we have a reasonable fit.

Example: Task Distraction Data

Example: Task Distraction Data

In an experimental study, the number of errors made in performing a specified task was recorded. The experiment investigated the influence of various predictors on the numbers of errors made.

There are two factor predictors (A, B) and one continuous covariate (X).

We have a balanced design with 15 people (replicates) in each factor-level subgroup.

Example: Task Distraction Data

Example: Task Distraction Data

A	Group	1 : Non-smoker
		2 : Delayed smoker
		3 : Active smoker
B	Task	1 : Pattern Recognition
		2 : Cognitive Task
		3 : Driving Simulation
X	Distraction Level	

Example: Task Distraction Data

We compare four models with the complete model

Complete Model : $A * B * X$

$$A + B + X + A:B + A:X + B:X + A:B:X$$

Number of parameters

Term	Parameters		Tot.
A	$(a-1)$	$= 3-1$	2
B	$(b-1)$	$= 3-1$	2
X	(1)		1
$A:B$	$(a-1)(b-1)$	$= 2 \times 2$	4
$A:X$	$(a-1)(1)$	$= 2 \times 1$	2
$B:X$	$(b-1)(1)$	$= 2 \times 1$	2
$A:B:X$	$(a-1)(b-1)(1)$	$= 2 \times 2 \times 1$	4
Total			17

Example: Task Distraction Data

We consider the following sequence of models:

- Reduced Model 1: M_{R_1}

$$A + B + X + A : X + B : X$$

- Reduced Model 2: M_{R_2}

$$A + B + X + B : X$$

- Reduced Model 3: M_{R_3}

$$B + X + B : X$$

- Reduced Model 4: M_{R_4}

$$B + X$$

Task Distraction Data: ANOVA F-Tests

$$
\begin{aligned}
M_C &: A + B + X + A:B + A:X + B:X + A:B:X \\
M_{R_1} &: A + B + X + A:X + B:X \\
M_{R_2} &: A + B + X + B:X \\
M_{R_3} &: B + X + B:X \\
M_{R_4} &: B + X
\end{aligned}
$$

C	R	SSE_C	SSE_R	k	g	F	F_α
M_C	M_{R_1}	5660.010	7627.479	17	9	5.084	2.02
M_{R_1}	M_{R_2}	7627.479	7971.274	9	7	2.817	3.07
M_{R_2}	M_{R_3}	7971.274	8404.654	7	5	3.452	3.07
M_{R_3}	M_{R_4}	8404.654	11154.715	5	3	21.105	3.07

Conclusions

Taking the comparisons in order:

1. M_C *vs* M_{R_1} : $F > F_\alpha$. This result is significant: Model M_{R_1} is not an adequate simplification of Model M_C

2. M_{R_1} *vs* M_{R_2} : $F < F_\alpha$. This result is not significant: Model M_{R_2} is an adequate simplification of Model M_{R_1}

3. M_{R_2} *vs* M_{R_3} : $F > F_\alpha$. This result is significant: Model M_{R_3} is not an adequate simplification of Model M_{R_2}

4. M_{R_3} *vs* M_{R_4} : $F > F_\alpha$. This result is significant: Model M_{R_4} is not an adequate simplification of Model M_{R_3}

Follow-up Analysis

In a follow-up analysis, it transpires that the model

$$A + B + X + A : B + A : X + B : X$$

ie selected.

Note: $R^2 = 0.863$ and $R^2_{adj} = 0.831$, so we have a good fit.

Note: we must take great care with the sequence of models.

Pitfalls of Regression Modelling

Five issues to bear in mind in ANOVA, Regression and General Linear Modelling.

1. Model assumptions
2. Data transformations
3. Model selection
4. Multicollinearity
5. Predicting beyond the range of the covariates

See textbook, section 12.12.

Chapter 3
Non-Parametric Statistics

In this section

- Textbook Chapters 13 & 14
- Chi-squared goodness of fit tests
- Non-parametric tests based on ranks
- One sample tests
- Two sample tests
- Permutation procedures

Non-Parametric Statistics

The previous statistical analysis methods studied (such as t-tests, ANOVA, Regression, Linear Modelling) have depended heavily on

distributional assumptions

specifically, that the data are Normally distributed.

We now seek statistical procedures that do not rely on this strong assumption. We term these methods

non-parametric or distribution-free

approaches.

Distribution-free tests for Categorical Data

Categorical data are data in which experimental units are allocated to one of a number of categories according to their characteristics. The categories are defined by one or more factors

- Treatment A, Treatment B - two categories
- Smoker/Former Smoker/Non Smoker - three categories.

Distribution-free tests for Categorical Data

In MATH 203: for 2×2 tables

- Two sample comparison of proportions
- Uses large sample approximation

See textbook, section 9.4.

Doll and Hill Data

Table 13.11. Smokers and non-smokers among male cancer patients and controls (Doll and Hill 1950)

	Smokers	Non-smokers	Total
Lung cancer	647	2	649
Controls	622	27	649

Juvenile Delinquency and Spectacle-Wearing

Table 10.14 Spectacle wearing among juvenile delinquen and non-delinquents who failed a vision test (Weindling *al* 1986)

		Juvenile delinquents	Non delinquents	Total
Spectacle wearers	Yes	1	5	6
	No	8	2	10
	Tota	9	7	16

Count data

The data are counts of experimental units that fall into each category. Suppose

- There are n experimental units in the study
- There are k categories
- The probabilities of the k outcomes are $p_1, \ldots, p_k$, where

$$p_1 + \cdots + p_k = 1$$

- The experimental units are independent
- The counts in the k categories are $n_1, \ldots, n_k$, where

$$n_1 + \cdots + n_k = n$$

Count data

The experimental design is termed a Multinomial Experiment

Note: The categories can arise as combinations of factor levels; we can have

- one-way classification (categories of a single factor, A)
- two-way classification (defined by combinations of levels of two factors, A and B)

and so on. The counts table is often called a contingency table and the entries in the table are called cells.

Count data

The idea can be extended to larger numbers of factors

$$A, B, C, \ldots$$

to produce a multi-way table.

We will focus on two-way tables, with r rows and c columns, yielding an $r \times c$ table.

Recap: The Binomial Distribution

Suppose random variable X can take values on the set

$$\{0, 1, \ldots, n\}$$

for some whole number $n \geqslant 1$. The Binomial distribution is a probability distribution for which

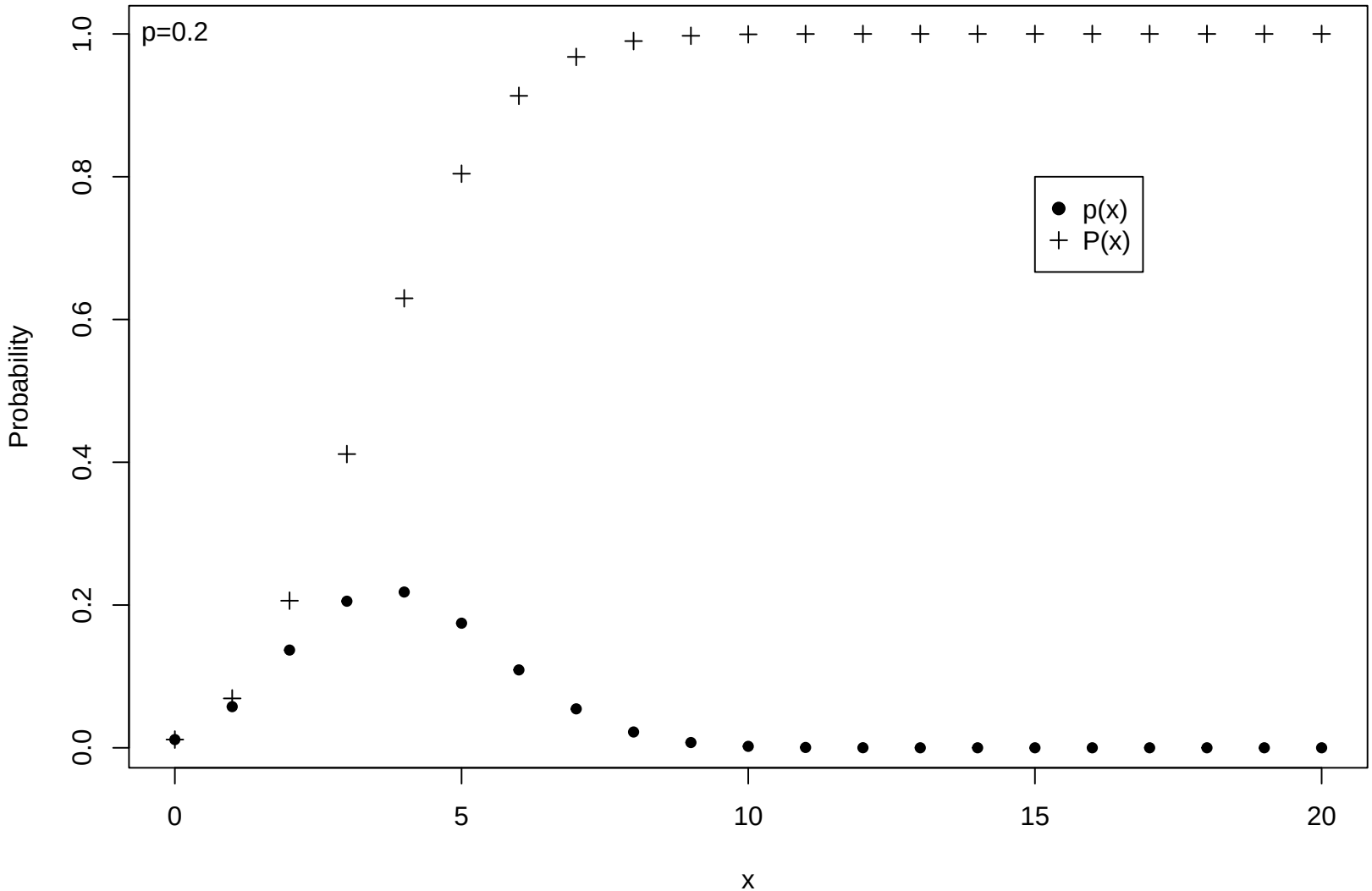
$$p(x) = \mathrm{P}[X = x] = \binom{n}{x} p^x (1-p)^{n-x} \qquad 0 \leqslant x \leqslant n$$

for some probability p.

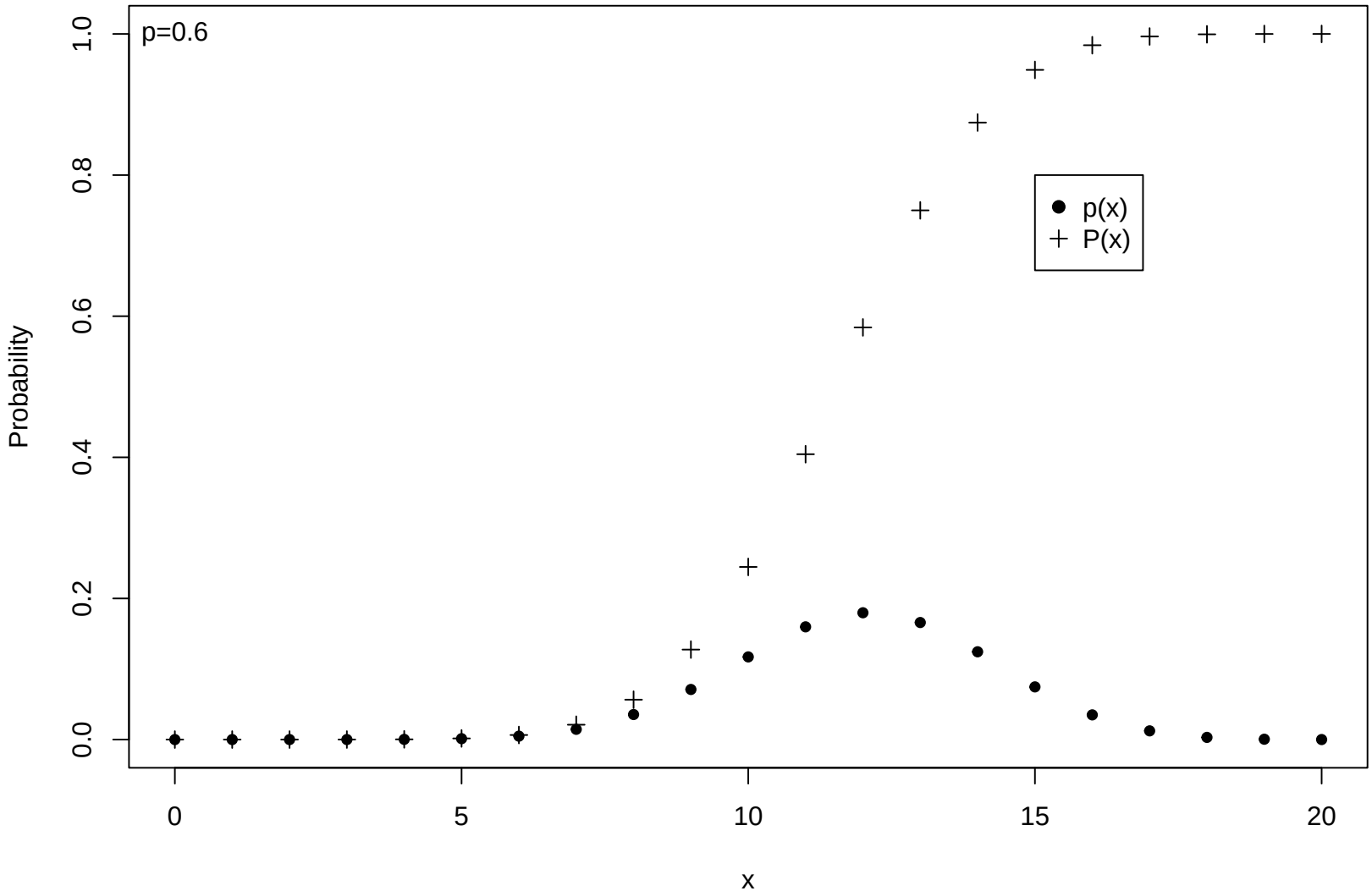
The cumulative probability function, $P(x)$, is defined by

$$P(x) = \mathrm{P}[X \leqslant x] = \sum_{t=0}^{x} \binom{n}{t} p^t (1-p)^{n-t}$$

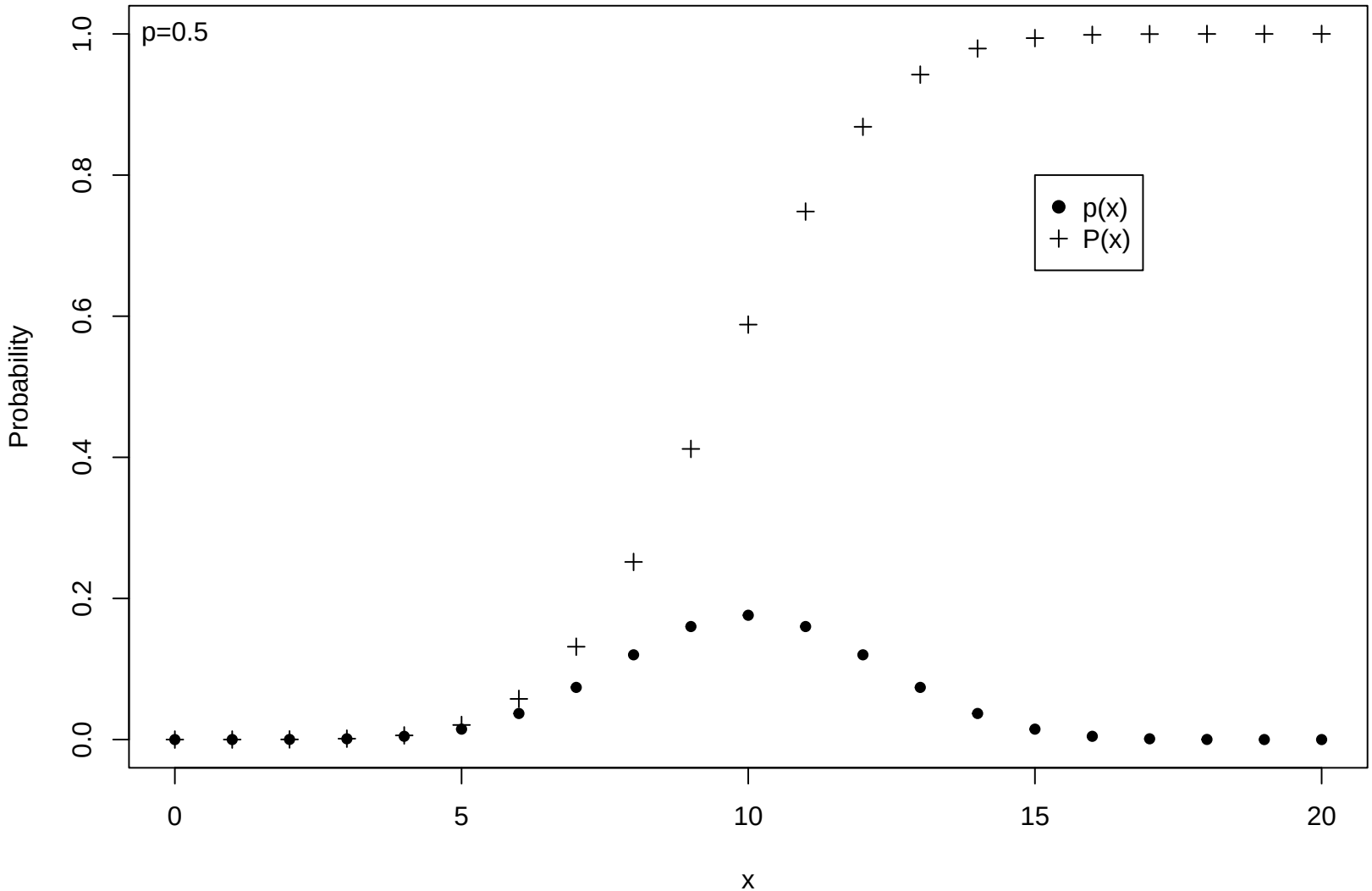
Recap: The Binomial Distribution



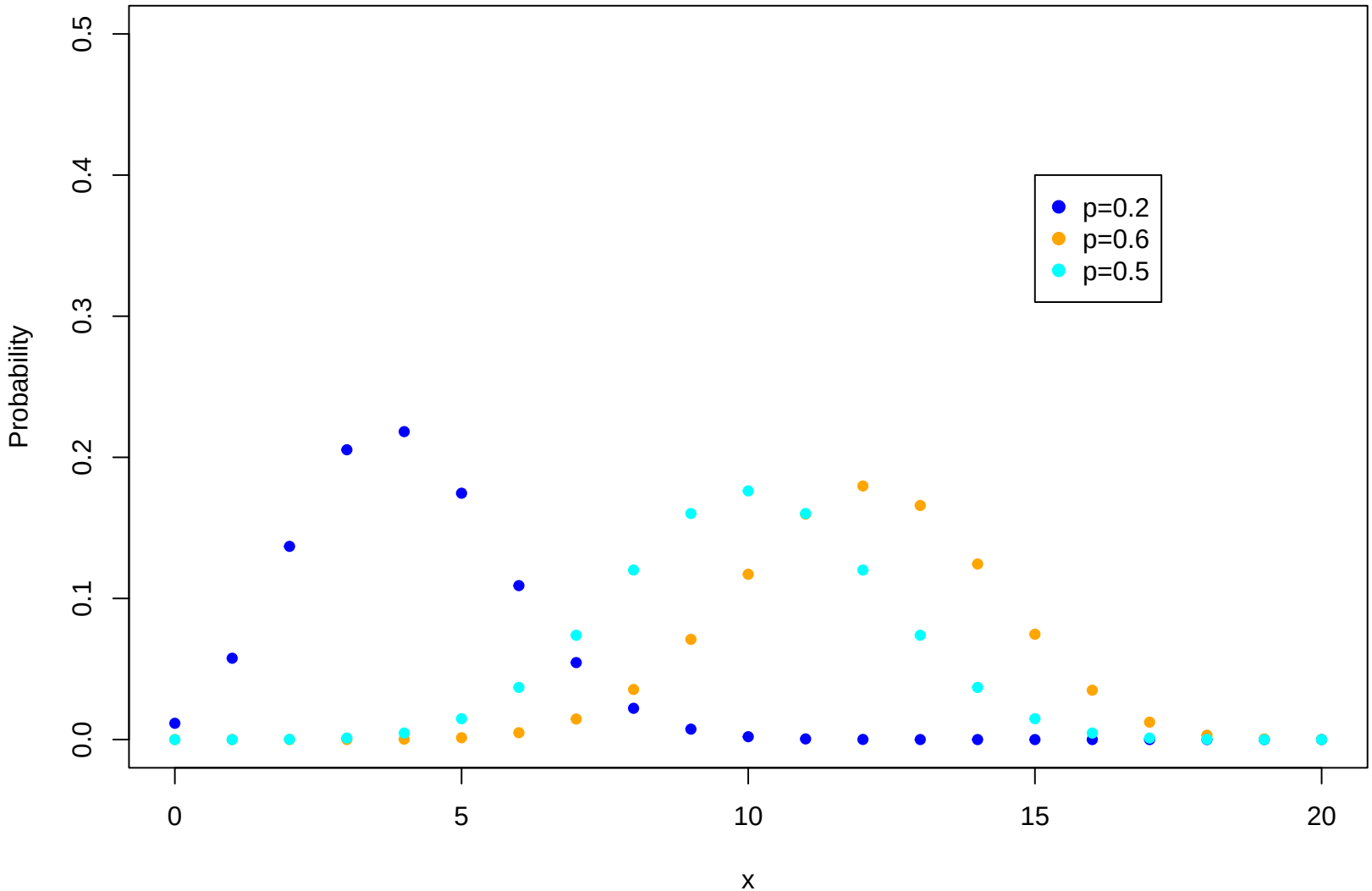
Recap: The Binomial Distribution



Recap: The Binomial Distribution



Recap: The Binomial Distribution



Tests for Binomial/Multinomial Data

What kinds of tests can be carried out for such data ?

1. Tests about $p_1, \ldots, p_k$
   - $H_0 : p_1 = \cdots = p_k = 1/k$
   - $H_0 : p_1, \ldots, p_k$ determined by distribution (eg Normal)
2. Tests about the factors A and B
   - are A and B dependent ?
   - i.e. does classification by A influence classification by B.

Chi-Squared Test for one-way tables

Suppose the null hypothesis completely specifies $p_1, \ldots, p_k$, that is, we have H_0 of the form

$$H_0 \ : \ p_1 = p_1^{(0)}, \ldots, p_k = p_k^{(0)}$$

where $p_1^{(0)}, \ldots, p_k^{(0)}$ are specified probabilities.

For example, for $k = 3$,

$$H_0 \ : \ p_1 = p_2 = p_3 = 1/3$$

or

$$H_0 \ : \ p_1 = 1/2, p_2 = p_3 = 1/4$$

Chi-Squared Test for one-way tables

To test this hypothesis against the alternative hypothesis

$$H_a \ : \ H_0 \text{ not true.}$$

we use the test statistic

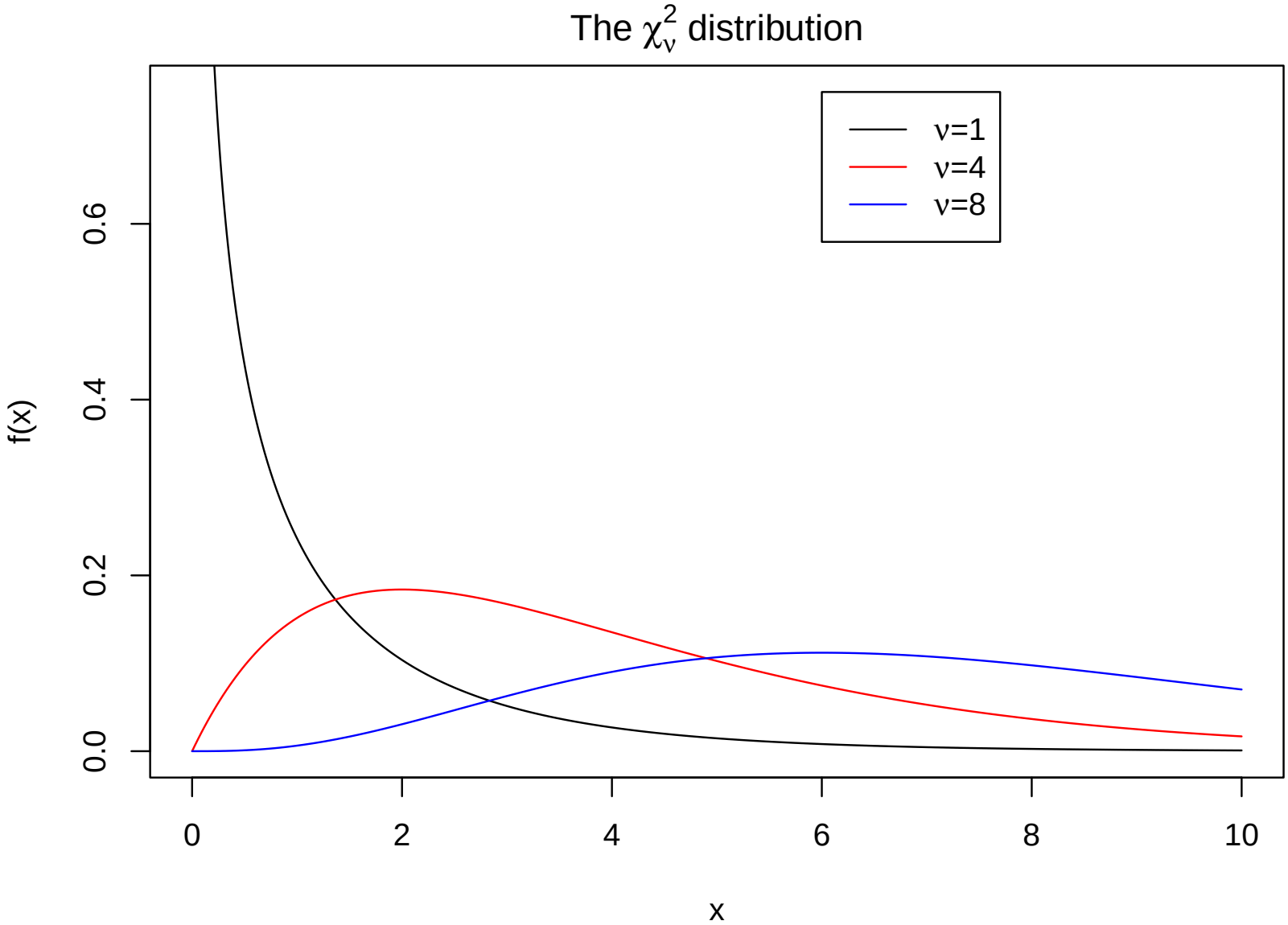
$$\mathrm{X}^2 = \sum_{i=1}^{k} \frac{\left(n_i - np_i^{(0)}\right)^2}{np_i^{(0)}}$$

If H_0 is true,

$$\mathrm{X}^2 \approx \chi^2_{k-1}.$$

That is, X^2 is approximately distributed as a chi-squared distribution with $k-1$ degrees of freedom.

Chi-Squared Test for one-way tables



Chi-Squared Test for one-way tables

Table: Upper a points of the χ^2_ν distribution

ν	a				ν	a			
	0.100	0.050	0.010	0.001		0.100	0.050	0.010	0.001
1	2.706	3.841	6.635	10.828	11	17.275	19.675	24.725	31.264
2	4.605	5.991	9.210	13.816	12	18.549	21.026	26.217	32.909
3	6.251	7.815	11.345	16.266	13	19.812	22.362	27.688	34.528
4	7.779	9.488	13.277	18.467	14	21.064	23.685	29.141	36.123
5	9.236	11.070	15.086	20.515	15	22.307	24.996	30.578	37.697
6	10.645	12.592	16.812	22.458	16	23.542	26.296	32.000	39.252
7	12.017	14.067	18.475	24.322	17	24.769	27.587	33.409	40.790
8	13.362	15.507	20.090	26.124	18	25.989	28.869	34.805	42.312
9	14.684	16.919	21.666	27.877	19	27.204	30.144	36.191	43.820
10	15.987	18.307	23.209	29.588	20	28.412	31.410	37.566	45.315

Chi-Squared Test for one-way tables

In this formula

- n_i is the observed count in cell i
- $np_i^{(0)}$ is the expected count in cell i if H_0 is true.

Sometimes the formula is written

$$\mathrm{X}^2 = \sum_{i=1}^{k} \frac{(O_i - E_i)^2}{E_i}$$

where O_i is the observed count, and E_i is the expected count assuming H_0 is true.

Chi-Squared Test for one-way tables

If

$$\mathrm{X}^2 > \chi^2_{\alpha,k-1}$$

then we reject H_0 at the α significance level, where

$$\chi^2_{\alpha,k-1}$$

is the upper α point of the chi-squared distribution with $k-1$ degrees of freedom.

Two-way tables

For the two-way table, we can test the hypothesis

H_0 : Factor A and Factor B levels are assigned independently

that is, classification by Factor A is independent of classification by Factor B.

Two-way tables

We use the same test statistic that can be rewritten

$$\mathrm{X}^2 = \sum_{i=1}^{r}\sum_{j=1}^{c} \frac{\left(n_{ij} - \widehat{n}_{ij}\right)^2}{\widehat{n}_{ij}}$$

where

$$\widehat{n}_{ij} = \frac{n_{i.}n_{.j}}{n} \qquad n_{i.} = \sum_{j=1}^{c} n_{ij} \qquad n_{.j} = \sum_{i=1}^{r} n_{ij}.$$

The terms $n_{i.}$ and $n_{.j}$ are the row and column totals for row i and column j respectively.

Two-way tables

If H_0 is true

$$\mathrm{X}^2 \approx \chi^2_{(r-1)(c-1)}$$

i.e. the degrees of freedom quantity is $(r-1)(c-1)$.

The test proceeds as before: we check whether

$$\mathrm{X}^2 > \chi^2_{\alpha,(r-1)(c-1)}$$

and if so, we reject H_0.

Example: 2×3 contingency table

Patients suffering from a chronic condition are recruited to a randomized clinical trial with three treatment groups:

- Drug 1
- Drug 2
- Drug 3

The alleviation of the chronic condition was studied:

- Condition did not improve
- Condition did improve

Data:

Improved	Drug 1	2	3	Total
No	10	12	10	32
Yes	10	18	15	43
Total	20	30	25	75

Is there a difference between the improvement rates across drugs ?

Notation:

Improved	Drug 1	2	3	Total
No	n_{11}	n_{12}	n_{13}	$n_{1.}$
Yes	n_{21}	n_{22}	n_{23}	$n_{2.}$
Total	$n_{.1}$	$n_{.2}$	$n_{.3}$	n

Scientific question

Null Hypothesis:

$$H_0 \quad : \quad \text{Improvement rate is the same across drugs.}$$

If H_0 is true, there is no difference between the drugs

- the elements in the columns of the table are realizations of three independent Binomial sampling schemes with the same probability of success p.

The best estimate of the common improvement rate p is

$$\widehat{p} = \frac{n_{2.}}{n} = \frac{43}{75} = 0.573$$

Scientific question

That is

$$\begin{aligned} n_{21} &\sim Binomial(n_{.1}, p) \\ n_{22} &\sim Binomial(n_{.2}, p) \\ n_{23} &\sim Binomial(n_{.3}, p) \end{aligned}$$

and thus, under H_0, we use the fitted values determined by the expectations of these distributions

$$\begin{aligned} \widehat{n}_{21} &= n_{.1}\widehat{p} = \frac{n_{.1}n_{2.}}{n} \\ \widehat{n}_{22} &= n_{.2}\widehat{p} = \frac{n_{.2}n_{2.}}{n} \\ \widehat{n}_{23} &= n_{.3}\widehat{p} = \frac{n_{.3}n_{2.}}{n} \end{aligned}$$

Scientific question

We can complete the “fitted" or “expected" table by noting that we must have

$$\begin{aligned} n_{.1} &= \hat{n}_{11} + \hat{n}_{21} \\ n_{.2} &= \hat{n}_{12} + \hat{n}_{22} \\ n_{.3} &= \hat{n}_{13} + \hat{n}_{23} \end{aligned}$$

This is sufficient to allow us to complete the “fitted" table

Fitted table under H_0

Improved	Drug 1	2	3	Total
No	8.533	12.800	10.667	32
Yes	11.467	17.200	14.333	43
Total	20.000	30.000	25.000	75

so that the test statistic is

$$\mathrm{X}^2_{obs} = \sum_{i=1}^{2}\sum_{j=1}^{3} \frac{\left(n_{ij} - \widehat{n}_{ij}\right)^2}{\widehat{n}_{ij}} = 0.599$$

Completing the test

Under H_0, we have that

$$\mathrm{X}^2 \approx \chi^2_{(r-1)(c-1)}$$

Here $r = 2$, $c = 3$, so

$$\mathrm{X}^2 \approx \chi^2_2$$

From the textbook tables Table VII, we find that

$$\chi^2_{0.05,2} = 5.991$$

so we do not reject the null hypothesis, and conclude that there is no evidence of a difference between the improvement rates for the three drugs.

Example: Nose bleeds and Alcohol

Example: Nose bleeds and alcohol consumption

140 patients suffering from nose bleeds, and 113 controls were interviewed about their alcohol consumption, specifically whether they were

1. Non-drinkers
2. Occasional drinkers
3. Regular drinkers

Scientific question

Is there a difference between the proportions of different consumption groups in nose-bleed cases and controls ?

Example: Nose bleeds and Alcohol

	Group			
Improved	1	2	3	Total
Nose-bleed	47	30	63	140
Control	39	40	34	113
Total	86	70	97	253

Example: Nose bleeds and Alcohol

Using the `chisq.test` function:

```
> NoseBleeds<-matrix(c(47,30,63,39,40,3), nrow=2,byrow=T)
> chisq.test(NoseBleeds)


+
+  Pearson's Chi-squared test
+
+ data:  NoseBleeds
+ X-squared = 44.61, df = 2, p-value = 2.056e-10
```

Example: Nose bleeds and Alcohol

By direct calculation:

```
> #Form the fitted values from the row and column totals
> NB.rowtots<-apply(NoseBleeds,1,sum)
> NB.coltots<-apply(NoseBleeds,2,sum)
> NB.fitted<-outer(NB.rowtots,NB.coltots,"*")/sum(NoseBleeds)
> NB.fitted

+          [,1]     [,2]     [,3]
+ [1,] 54.23423 44.14414 41.62162
+ [2,] 31.76577 25.85586 24.37838

> (Xsq<-sum((NoseBleeds-NB.fitted)^2/NB.fitted))

+ [1] 44.61003

> (pval<-1-pchisq(Xsq,2))

+ [1] 2.056156e-10
```

Validity of the chi-squared test

For the Chi-squared test to be valid, we need the expected cell counts

$$np_i^{(0)} \qquad i = 1, \ldots, k$$

or

$$\widehat{n}_{ij} \qquad i = 1, \ldots, r, j = 1, \ldots, c$$

to be sufficiently large. The convention is to require the expected value to be greater than five.

An Exact Test

If $r = c = 2$ we have a 2×2 table, and another exact test can be used: this does not rely on the large sample approximation.

Fisher's Exact Test

- another test for independence of assignment of the row and column factor levels
- test statistic and null distribution are complicated (based on the hypergeometric distribution)
- R computes test statistic and p-value.

An Exact Test

The test statistic, T, is the (1,1) cell count used to test

$$H_0 \ : \ \text{Factors are assigned independently.}$$

If H_0 is true, then

$$T \sim Hypergeometric\left(n_{1.}, n_{2.}, n_{.1}, n_{.2}\right)$$

This is a test carried out conditional on the row and column totals.

An Exact Test

Example: Juvenile Delinquency and Spectacle Wearing

Is there any association between the two factors ?

A : Spectacle Wearing (Yes/No)

B : Juvenile Delinquent (Yes/No)

		Delinquent		
		Yes	No	$n_{i.}$
Spectacles	Yes	1	5	6
	No	8	2	10
	$n_{.j}$	9	7	16

An Exact Test

```
> JD<-matrix(c(1,8,5,2),2,2)
> chisq.test(JD, correct=FALSE)

+ Warning in chisq.test(JD, correct = FALSE): Chi-squared approximation may be incorrect


+
+  Pearson's Chi-squared test
+
+ data:  JD
+ X-squared = 6.1122, df = 1, p-value = 0.01343

> fisher.test(JD)


+
+  Fisher's Exact Test for Count Data
+
+ data:  JD
+ p-value = 0.03497
+ alternative hypothesis: true odds ratio is not equal to 1
+ 95 percent confidence interval:
+  0.0009525702 0.9912282442
+ sample estimates:
+ odds ratio
+ 0.06464255
```

An Exact Test

```
> fisher.test(JD, alternative='less')


+
+  Fisher's Exact Test for Count Data
+
+ data:  JD
+ p-value = 0.02448
+ alternative hypothesis: true odds ratio is less than 1
+ 95 percent confidence interval:
+  0.0000000 0.7290771
+ sample estimates:
+ odds ratio
+ 0.06464255
```

An Exact Test

Here we have $T_{obs} = 1$. The possible values of T are

$$\{\max(0, n_{1.} - n_{.2}), \ldots, \min(n_{1.}, n_{.1})\}$$

which here corresponds to the set

$$\{0, 1, 2, 3, 4, 5, 6\}$$

as

$$n_{.1} = 9 \quad n_{.2} = 7 \quad n_{1.} = 6 \quad n_{.2} = 10.$$

An Exact Test

In this Example

$$\begin{aligned} \text{One-sided p-value} &: P_0(T \leqslant 1) \\ \text{Two-sided p-value} &: P_0(T \leqslant 1) + P_0(T \geqslant 6) \end{aligned}$$

That is

$$0.0244755 \quad \text{and} \quad 0.034965$$

respectively.

An Exact Test

Example: Juvenile Delinquency and Spectacle Wearing

Chi-squared Test:

$$\chi^2_{obs} = 6.112$$

Compare with $\chi^2_{(r-1)(c-1)} = \chi^2_1$; we have

$$\chi^2_{0.05,1} = 3.841$$

and a p-value of 0.013. Therefore we reject H_0.

Fisher's Exact Test: p-value is 0.035 (2-sided) or 0.024 (1-sided).

Thus we reject H_0 and we have evidence of association between the factors.

Case-Control Studies

A case-control study is an observational study where we collect and assess

- a sample of cases (disease sufferers)
- a sample of controls (healthy patients)

We investigate the possible association between disease status and a factor that takes two levels. A 2×2 table of counts is formed for all combinations of disease status/factor level.

Case-Control Studies

Example: BCG Vaccination and Leprosy

In this example: we assess the possibility of an association between the two factors.

Disease Status : Leprosy Sufferer (Yes/No)

Factor : Vaccination Scar (Yes/No)

		Disease Status: Case (Yes)	Disease Status: Control (No)	$n_{i.}$
Scar	Yes	101	554	655
	No	159	446	605
	$n_{.j}$	260	1000	1260

Odds Ratio

An alternative test statistic is based on the odds ratio

$$\text{O.R.} = \frac{n_{11}n_{22}}{n_{12}n_{21}} = \widehat{\psi}$$

say. The test statistic is

$$Z = \frac{\log \widehat{\psi}}{\text{e.s.e.}(\log \widehat{\psi})}$$

where

$$\text{e.s.e.}(\log \widehat{\psi}) = \sqrt{\frac{1}{n_{11}} + \frac{1}{n_{12}} + \frac{1}{n_{21}} + \frac{1}{n_{22}}}$$

Odds Ratio

That is,

$$Z = \frac{\log n_{11} + \log n_{22} - \log n_{12} - \log n_{21}}{\sqrt{\frac{1}{n_{11}} + \frac{1}{n_{12}} + \frac{1}{n_{21}} + \frac{1}{n_{22}}}}$$

If H_0

H_0 : No association between factor and disease status

is true, it follows that

$$Z \dot{\sim} Normal(0, 1)$$

Here $\log$ means $\ln$ or natural log.

Odds Ratio

Example: BCG Vaccination and Leprosy

$$n_{11} = 101, n_{12} = 554, n_{21} = 159, n_{22} = 446$$

Therefore

$$\widehat{\psi} = \frac{n_{11}n_{22}}{n_{12}n_{21}} = 0.511 \qquad \log\widehat{\psi} = -0.671$$

and

$$\text{e.s.e.}(\log\widehat{\psi}) = \sqrt{\frac{1}{n_{11}} + \frac{1}{n_{12}} + \frac{1}{n_{21}} + \frac{1}{n_{22}}} = 0.142$$

Odds Ratio

Example: BCG Vaccination and Leprosy

Therefore

$$Z_{obs} = \frac{-0.671}{0.142} = -4.717$$

For a test at $\alpha = 0.05$, the two-sided critical values are ± 1.96, so we Reject H_0.

Odds Ratio

Example: Smoking and Lung Cancer

Here $n_{11} = 647, n_{12} = 622, n_{21} = 2, n_{22} = 27$, so

$$\log \widehat{\psi} = \log \frac{647 \times 27}{2 \times 622} = 2.642$$

and

$$\text{e.s.e.}(\log \widehat{\psi}) = \sqrt{\frac{1}{647} + \frac{1}{2} + \frac{1}{622} + \frac{1}{27}} = 0.735$$

so

$$Z_{obs} = \frac{2.642}{0.735} = 3.590$$

Odds Ratio

Example: Smoking and Lung Cancer

For a test at $\alpha = 0.05$, the two-sided critical value is ± 1.96, so we

$$\text{Reject } H_0$$

and report evidence for association.

Rank Tests

We seek non-parametric or distribution-free tests for hypotheses relating to single samples, the equivalents of one-sample z- or t-tests, which rely on the normality of the samples.

Normally these tests are formulated in terms of ranks of the data to give

Rank Tests

Rank Tests

For example, if the data are

$$0.24 \quad 3.16 \quad 1.97 \quad 2.10 \quad 0.92$$

we sort them into ascending order, and assign ranks in order

	0.24	0.92	1.97	2.10	3.16
Rank	1	2	3	4	5

The tests depend on the behaviour of statistics computed in terms of the ranks.

Rank Tests

The statistical tasks are the same as in any other hypothesis testing setting, that is we need to

- define the null hypothesis
- define a suitable test statistic
- compute the null distribution
- define and utilize the rejection region to complete the test.

Single Population Tests

Rather than test the mean, we typically will test the median, x_{MED}, where

$$\Pr[\text{Observation} \leqslant x_{\text{MED}}] = \frac{1}{2}$$

the halfway point of the distribution.

The sample median is the halfway point of the sorted sample.

Single Population Tests

In a one-sample problem, we observe data

$$x_1, \ldots, x_n$$

and we wish to carry out tests related to the underlying population.

If η is the population median, we wish to test, for example,

$$H_0 \ : \ \eta = \eta_0$$

for some fixed value η_0.

Two Population Tests

In a two-sample problem, we observe data

$$x_{11}, \ldots, x_{n1} \qquad x_{12}, \ldots, x_{n2}$$

from two populations.

We seek a non-parametric equivalent to the two-sample t-test, but instead of testing population means,

$$H_0 \ : \ \mu_1 = \mu_2$$

we test population medians

$$H_0 \ : \ \eta_1 = \eta_2$$

One- and Two- sample tests

- In the one sample case we use the

  sign test

  to test hypotheses about η

- In the two sample case we use the

  Wilcoxon Rank Sum or Mann-Whitney U test.

One- and Two- sample tests

For the MWW test

- Label the samples so that $n_1 > n_2$ (i.e. sample 1 is the one with the larger sample size)
- The test is based on the **sum of the ranks** for the data from sample 2. We
  - Pool and sort the data, retaining the sample labels
  - Assign ranks in ascending order
  - The rank sum for sample 1 is R_1, and the rank sum for sample 2 is denoted R_2.

Example: Mann-Whitney-Wilcoxon test statistic

SAMPLE 1	$n_1 = 4$	0.31	0.48	1.02	3.11
SAMPLE 2	$n_2 = 3$	0.16	0.20	1.97	

yields the following ranked data

SAMPLE	2	2	**1**	**1**	**1**	2	**1**
	0.16	0.20	**0.31**	**0.48**	**1.02**	1.97	**3.11**
RANK	1	2	**3**	**4**	**5**	6	**7**

Thus the rank sum for sample 1 and sample 2 are

$$R_1 = 3 + 4 + 5 + 7 = 19 \qquad R_2 = 1 + 2 + 6 = 9.$$

One- and Two- sample tests

Two related test statistics can be used

- Wilcoxon Rank Sum Statistic

$$W = R_2$$

- Mann-Whitney U Statistic

$$U = R_2 - \frac{n_2(n_2 + 1)}{2}$$

One- and Two- sample tests

- Large Sample Test If $n_2 \geqslant 10$, a large sample test based on the Z statistic

$$Z = \frac{U - \frac{n_1 n_2}{2}}{\sqrt{\frac{n_1 n_2 (n_1 + n_2 + 1)}{12}}}$$

can be used. Under $H_0 \ : \ \eta_1 = \eta_2$,

$$Z \dot{\sim} \text{Normal}(0, 1)$$

and the test at $\alpha = 0.05$ uses the following critical values

$$\begin{aligned} H_a \ &: \ \eta_1 > \eta_2 \quad \text{then} \quad C_R = -1.645 \\ H_a \ &: \ \eta_1 < \eta_2 \quad \text{then} \quad C_R = 1.645 \\ H_a \ &: \ \eta_1 \neq \eta_2 \quad \text{then} \quad C_R = \pm 1.960 \end{aligned}$$

One- and Two- sample tests

- Small Sample Test If $n_1 < 10$, an exact but more complicated test can be used.
  - The test statistic is R_2.
  - The null distribution under the hypothesis $H_0 : \eta_1 = \eta_2$ can be computed, but it is complicated.

    The table in McClave and Sincich gives the critical values (T_L and T_U) that determine the rejection region for different n_1 and n_2 values up to 10.

One- and Two- sample tests

- One-sided tests:

$$H_a \ : \ \eta_1 > \eta_2 \quad \text{Rejection Region is} \quad R_2 \leqslant T_L$$
$$H_a \ : \ \eta_1 < \eta_2 \quad \text{Rejection Region is} \quad R_2 \geqslant T_U$$

  These are tests at the $\alpha = 0.025$ significance level.

- Two-sided tests:

$$H_a \ : \ \eta_1 \neq \eta_2$$

  Rejection Region is

$$\{R_2 \leqslant T_L \text{ or } R_2 \geqslant T_U\}$$

  This is a test at the $\alpha = 0.05$ significance level.

One- and Two- sample tests

Other two sample tests are available:

- Kolmogorov-Smirnov Test
- Moses Extreme Reactions Test
- Wald-Wolfowitz Runs Test

None make distributional assumptions, all perform best when the sample size is large.

Comparing Two Dependent Samples

Suppose we have repeat measurements on the same experimental units.

In this case, the within-subject data are dependent; we have pairing of observations.

We can use the

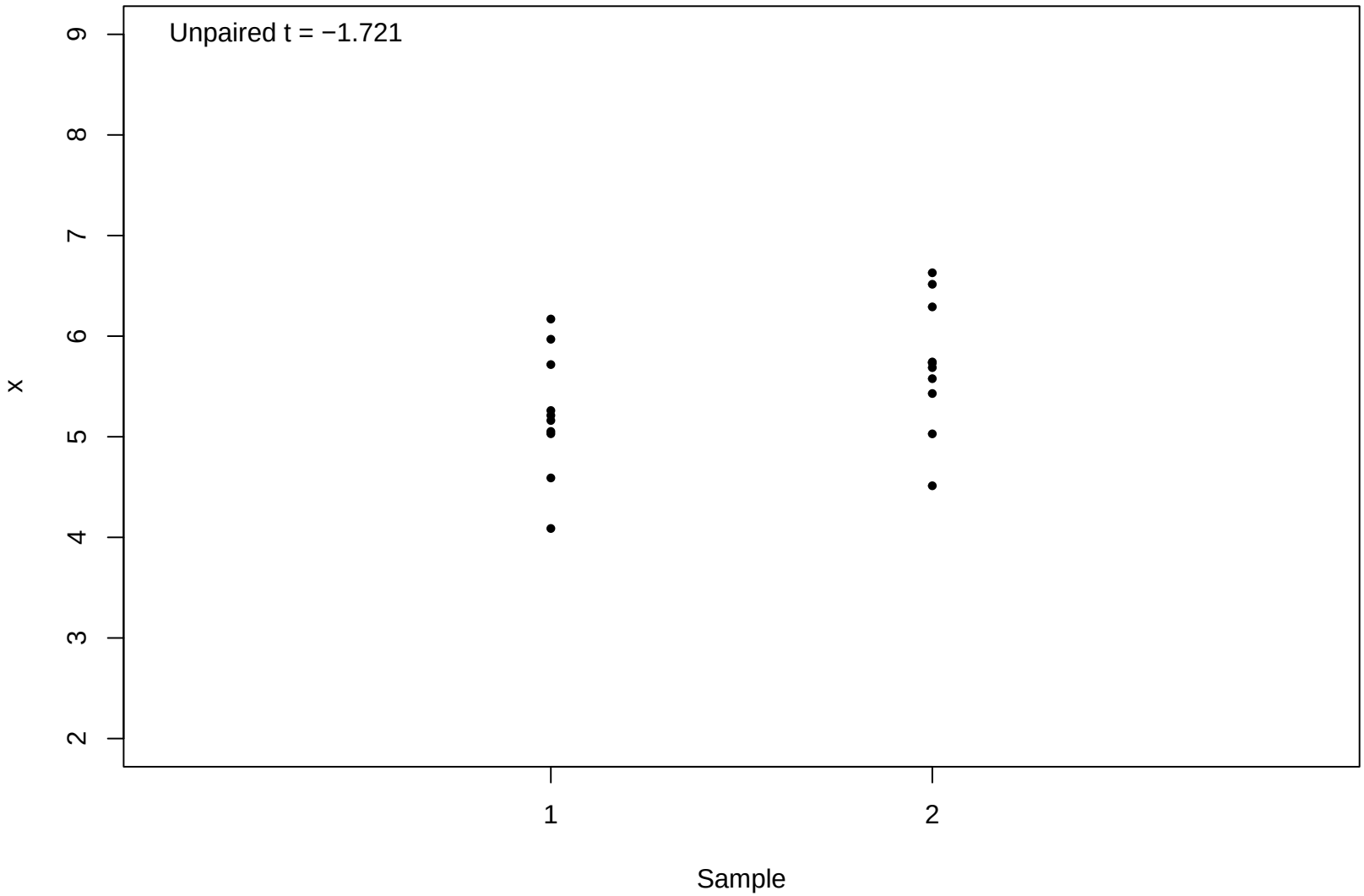
Wilcoxon Signed Rank Test

The Wilcoxon Signed Rank Test

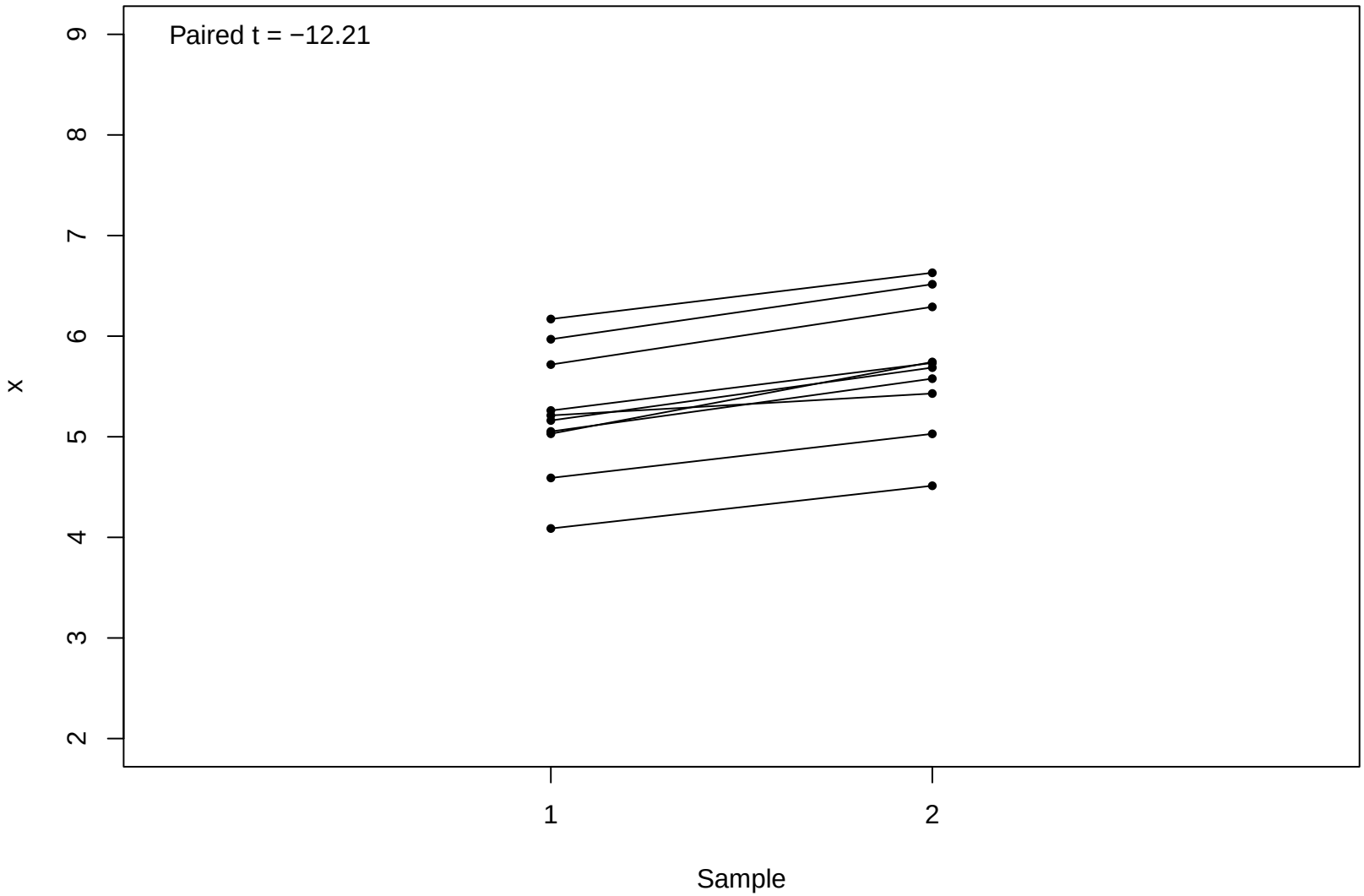
Remember:

- The Wilcoxon Test: a test for independent samples.
- The Wilcoxon Signed Rank Test: a test for paired data

Two sample test without Pairing



Two sample test with Pairing



The Wilcoxon Signed Rank Procedure

For paired data

- pre-treatment $(y_{11}, \ldots, y_{n1})$
- post-treatment: $(y_{12}, \ldots, y_{n2})$ data,

the procedure is as follows:

1. Form the differences

$$x_i = y_{i1} - y_{i2} \qquad i = 1, \ldots, n.$$

2. Form the absolute values $s_i = |x_i|$, for $i = 1, \ldots, n$.
3. Drop every $x_i = 0$, leaving a sample size of m.

The Wilcoxon Signed Rank Procedure

4. Assign ranks $r_1, \ldots, r_m$ to $s_1, \ldots, s_m$ after sorting into ascending order.
5. Form

$$T_+ = \sum_{i=1}^{m} r_i z_i \qquad\qquad T_- = \sum_{i=1}^{m} r_i (1 - z_i)$$

where

$$z_i = \begin{cases} 1 & x_i > 0 \\ 0 & x_i < 0 \end{cases}$$

Computing the Null Distribution

The behaviour of T_+ and T_- can be predicted under the null hypothesis that the original data are generated from

continuous distributions

with the same median: we will never get $x_i = 0$.

- If K is the number of positive x_i values, then we have that $K \sim Binomial(n, 1/2)$, as in the sign test.
- Conditional on $K = k$, we can compute the distribution

$$P_0(T_+ = t | K = k)$$

  by studying the behaviour of the ranks.
- We then use the Theorem of Total Probability to compute

$$P_0(T_+ = t)$$

Comparing Three or More Populations

We now seek non-parametric equivalents to ANOVA useful for different designs. We study tests for

(a) the Completely Randomized Design (CRD)

(b) the Randomized Block Design (RBD)

For (a) we use the

Kruskal-Wallis Test

and for (b) we use the

Friedman Test.

Summary of the Non-Parametric Tests

- Chi-Squared Test : Goodness of Fit/independence in contingency tables
- Sign Test : One Sample (equivalent of one sample t-test)
- Mann-Whitney-Wilcoxon : Two Sample (equivalent of two sample t-test)
- Wilcoxon Signed Rank : Paired Data
- Kruskal-Wallis : one-way layout, multigroup comparison - equivalent of ANOVA for CRD.
- Friedman : two-way blocked layout, equivalent of two-way ANOVA for RBD.

Summary of the Non-Parametric Tests

Pros:

- No distributional assumptions
- Applicable for most sorts of data
- Large sample approximations make them easy to implement

Cons:

- Low power compared to parametric tests (i.e. often do not reject H_0 when they should - prone to Type II Error)
- Small sample null distributions difficult to compute

Rank Correlation

To measure the association between two variables, we previously used the correlation coefficient, r; for data $x_1, \ldots, x_n$ and $y_1, \ldots, y_n$,

$$r = \frac{SS_{xy}}{\sqrt{SS_{xx} SS_{yy}}}$$

where

$$SS_{xy} = \sum_{i=1}^{n} (x_i - \overline{x})(y_i - \overline{y}) \quad SS_{xx} = \sum_{i=1}^{n} (x_i - \overline{x})^2 \quad SS_{yy} = \sum_{i=1}^{n} (y_i - \overline{y})^2$$

r is a measure of the linear association between X and Y

Pearson Product Moment Coefficient of Correlation

Rank Correlation

A more general measure of association is the

Spearman Rank Correlation Coefficient

We compute this as follows:

1. For each sample separately, compute the ranks of the data. Let the ranks for the data $x_1, \ldots, x_n$ and $y_1, \ldots, y_n$ by $u_1, \ldots, u_n$ and $v_1, \ldots, v_n$ respectively.

2. Compute

$$r_S = \frac{SS_{uv}}{\sqrt{SS_{uu} SS_{vv}}}$$

   ie the Pearson correlation between the ranks.

r_S is the Spearman Correlation.

Rank Correlation

Notes:

1. If there are no ties in the data

$$r_S = 1 - \frac{6\sum_{i=1}^{n} d_i^2}{n(n^2-1)}$$

   where $d_i = u_i - v_i$.

2. r_S is potentially a measure of the non-linear association between X and Y.

   The calculation can be applied directly to rank data i.e. $u_1, \ldots, u_n$ and $v_1, \ldots, v_n$ can be preference ranks given by two observers.

Tests for r_S

To test

$$H_0 \ : \ \rho = 0$$

vs

$$\begin{aligned} (1) \quad H_a &: \quad \rho > 0 \\ (2) \quad H_a &: \quad \rho < 0 \\ (3) \quad H_a &: \quad \rho \neq 0 \end{aligned}$$

We may use r_S as a test statistic. The distribution of r_S under H_0 is tabulated in the textbook.

Tests for r_S

If Spearman_α is the α tail quantile of the null distribution, we have the following rejection regions:

$$
\begin{aligned}
(1) &: \text{Reject } H_0 \text{ if } r_S > \text{Spearman}_\alpha \\
(2) &: \text{Reject } H_0 \text{ if } r_S < -\text{Spearman}_\alpha \\
(3) &: \text{Reject } H_0 \text{ if } |r_S| > \text{Spearman}_{\alpha/2}
\end{aligned}
$$

See textbook Table XI.

Randomization and Permutation Tests

Randomization or Permutation procedures are useful for computing exact null distributions for non-parametric test statistics when sample sizes are small.

We focus first on two sample comparisons; suppose that two data samples $x_1 \ldots, x_{n_1}$ and $y_1 \ldots, y_{n_2}$ (where $n_1 \geqslant n_2$) have been obtained, and we wish to carry out a comparison of the two populations from which the samples are drawn.

The Wilcoxon test statistic, W, is the sum of the ranks for the second sample.

Randomization and Permutation Tests

The permutation test proceeds as follows:

1. Let $n = n_1 + n_2$. Assuming that there are no ties, the pooled and ranked samples will have ranks

$$1 \quad 2 \quad 3 \quad \ldots \quad n$$

2. The test statistic is $W = R_2$, the rank sum for sample two items. For the observed data, W will be the sum of n_2 of the ranks given in the list above.

Randomization and Permutation Tests

3. If H_0

$$H_0 \ : \text{ No difference between popn. 1 and popn. 2}$$

were true, then we would expect no pattern in the arrangements of the group labels when sorted into ascending order. That is, the sorted data would give rise a random assortment of group 1 and group 2 labels.

Randomization and Permutation Tests

4. To obtain the exact distribution of W under H_0, we could compute W for all possible permutations of the group labels, and then form the probability distribution of the values of W.

   We call this the permutation null distribution.

5. But W is a rank sum, so we can compute the permutation null distribution simply by tabulating all possible subsets of size n_2 of the set of ranks $\{1, 2, 3, \ldots, n\}$.

6. There are

$$\binom{n}{n_2} = \frac{n!}{n_1!\, n_2!} = N$$

say possible subsets of size n_2. For example, for $n = 6$ and $n_2 = 2$, the number of subsets of size n_2 is

$$\binom{8}{2} = \frac{8!}{6!\, 2!} = 28$$

However, the number of subsets increases dramatically as n increases; for $n_1 = n_2 = 10$, so that $n = 20$, the number of subsets of size n_2 is

$$\binom{20}{10} = \frac{20!}{10!\, 10!} = 184756$$

7. The rejection region and p-value are computed from the permutation null distribution. Let $W_i, i = 1, \ldots, N$ denote the value of the Wilcoxon statistic for the N possible subsets of the ranks of size n_2.

   The probability that the test statistic, W, is less than or equal to w is

$$P_0(W \leqslant w) = \frac{\text{Number of } W_i \leqslant w}{N}$$

Randomization and Permutation Tests

We seek the values of w that give the appropriate rejection region, $\mathcal{R}$, so that

$$P_0(W \in \mathcal{R}) = \frac{\text{Number of } W_i \in \mathcal{R}}{N} = \alpha$$

It may not be possible to find critical values, and define $\mathcal{R}$, so that this probability is exactly α as the distribution of W is discrete.

Example: Simple Example

Suppose $n_1 = 7$ and $n_2 = 3$. There are

$$\binom{10}{3} = \frac{10!}{7!\,3!} = 120$$

subsets of the ranks $\{1, 2, 3, \ldots, 10\}$ of size 3. The subsets are listed below, together with the rank sums.

Ranks			W	Ranks			W	Ranks			W	Ranks			W
1	2	3	6	1	7	8	16	2	7	10	19	4	6	7	17
1	2	4	7	1	7	9	17	2	8	9	19	4	6	8	18
1	2	5	8	1	7	10	18	2	8	10	20	4	6	9	19
1	2	6	9	1	8	9	18	2	9	10	21	4	6	10	20
1	2	7	10	1	8	10	19	3	4	5	12	4	7	8	19
1	2	8	11	1	9	10	20	3	4	6	13	4	7	9	20
1	2	9	12	2	3	4	9	3	4	7	14	4	7	10	21
1	2	10	13	2	3	5	10	3	4	8	15	4	8	9	21
1	3	4	8	2	3	6	11	3	4	9	16	4	8	10	22
1	3	5	9	2	3	7	12	3	4	10	17	4	9	10	23
1	3	6	10	2	3	8	13	3	5	6	14	5	6	7	18
1	3	7	11	2	3	9	14	3	5	7	15	5	6	8	19
1	3	8	12	2	3	10	15	3	5	8	16	5	6	9	20
1	3	9	13	2	4	5	11	3	5	9	17	5	6	10	21
1	3	10	14	2	4	6	12	3	5	10	18	5	7	8	20
1	4	5	10	2	4	7	13	3	6	7	16	5	7	9	21
1	4	6	11	2	4	8	14	3	6	8	17	5	7	10	22
1	4	7	12	2	4	9	15	3	6	9	18	5	8	9	22
1	4	8	13	2	4	10	16	3	6	10	19	5	8	10	23
1	4	9	14	2	5	6	13	3	7	8	18	5	9	10	24
1	4	10	15	2	5	7	14	3	7	9	19	6	7	8	21
1	5	6	12	2	5	8	15	3	7	10	20	6	7	9	22
1	5	7	13	2	5	9	16	3	8	9	20	6	7	10	23
1	5	8	14	2	5	10	17	3	8	10	21	6	8	9	23
1	5	9	15	2	6	7	15	3	9	10	22	6	8	10	24
1	5	10	16	2	6	8	16	4	5	6	15	6	9	10	25
1	6	7	14	2	6	9	17	4	5	7	16	7	8	9	24
1	6	8	15	2	6	10	18	4	5	8	17	7	8	10	25
1	6	9	16	2	7	8	17	4	5	9	18	7	9	10	26
1	6	10	17	2	7	9	18	4	5	10	19	8	9	10	27

There are 22 possible rank sums, $\{6, 7, 8, \ldots, 25, 26, 27\}$; the number of times each sum is observed is displayed in the table, with the corresponding probabilities and cumulative probabilities.

W	6	7	8	9	10	11	12	13	14	15	16
Frequency	1	1	2	3	4	5	7	8	9	10	10
Prob.	0.008	0.008	0.017	0.025	0.033	0.042	0.058	0.067	0.075	0.083	0.083
Cumulative Prob.	0.008	0.017	0.033	0.058	0.092	0.133	0.192	0.258	0.333	0.417	0.500

W	17	18	19	20	21	22	23	24	25	26	27
Frequency	10	10	9	8	7	5	4	3	2	1	1
Prob.	0.083	0.083	0.075	0.067	0.058	0.042	0.033	0.025	0.017	0.008	0.008
Cumulative Prob.	0.583	0.667	0.742	0.808	0.867	0.908	0.942	0.967	0.983	0.992	1.000

Thus, for example, the probability that $W = 19$ is 0.075, with a frequency of 9 out of 120. From this table, we deduce that

$$P_0(8 \leqslant W \leqslant 25) = 0.983 - 0.033 = 0.950$$

implying that the two-sided rejection region for $\alpha = 0.05$ is the set $\mathcal{R} = \{6, 7, 26, 27\}$.

Placenta Permeability Data

Example: Placenta Permeability Data

Measurements of placenta permeability are made on two groups of subjects.

The data and their ranks for are displayed below:

Group	1	1	1	1	1	1	1	1	1	1	2	2	2	2	2
Obs.	0.73	0.80	0.83	1.04	1.38	1.45	1.46	1.64	1.89	1.91	0.74	0.88	0.9	1.15	1.21
Rank	1	3	4	7	10	11	12	13	14	15	2	5	6	8	9

Placenta Permeability Data

Example: Placenta Permeability Data

Thus the Wilcoxon statistic is

$$W = R_2 = 2 + 5 + 6 + 8 + 9 = 30$$

Now, here $n_1 = 10$ and $n_2 = 5$. There are

$$\binom{15}{5} = \frac{15!}{10!\,5!} = 3003$$

subsets of the ranks $\{1, 2, 3, \ldots, 15\}$ of size 5.

In the permutation null distribution, the possible values of W are $\{15, 16, \ldots, 64, 65\}$; the probabilities are given below.

Placenta Permeability Data

W	15	16	17	18	19	20	21	22	23	24	25	26	27
Frequency	1	1	2	3	5	7	10	13	18	23	30	36	45
Prob.	0.000	0.000	0.001	0.001	0.002	0.002	0.003	0.004	0.006	0.008	0.010	0.012	0.015
Cumulative Prob.	0.000	0.001	0.001	0.002	0.004	0.006	0.010	0.014	0.020	0.028	0.038	0.050	0.065

W	28	29	30	31	32	33	34	35	36	37	38	39	40
Frequency	53	63	72	83	92	103	111	121	127	134	137	141	141
Prob.	0.018	0.021	0.024	0.028	0.031	0.034	0.037	0.040	0.042	0.045	0.046	0.047	0.047
Cumulative Prob.	0.082	0.103	0.127	0.155	0.185	0.220	0.257	0.297	0.339	0.384	0.430	0.477	0.523

W	41	42	43	44	45	46	47	48	49	50	51	52	53
Frequency	141	137	134	127	121	111	103	92	83	72	63	53	45
Prob.	0.047	0.046	0.045	0.042	0.040	0.037	0.034	0.031	0.028	0.024	0.021	0.018	0.015
Cumulative Prob.	0.570	0.616	0.661	0.703	0.743	0.780	0.815	0.845	0.873	0.897	0.918	0.935	0.950

W	54	55	56	57	58	59	60	61	62	63	64	65
Frequency	36	30	23	18	13	10	7	5	3	2	1	1
Prob.	0.012	0.010	0.008	0.006	0.004	0.003	0.002	0.002	0.001	0.001	0.000	0.000
Cumulative Prob.	0.962	0.972	0.980	0.986	0.990	0.994	0.996	0.998	0.999	0.999	1.000	1.000

Placenta Permeability Data

Example: Placenta Permeability Data

By inspection of the table, we see that

$$P_0(25 \leqslant W \leqslant 55) = 0.972 - 0.038 = 0.934$$

and

$$P_0(24 \leqslant W \leqslant 56) = 0.980 - 0.028 = 0.952$$

Placenta Permeability Data

Example: Placenta Permeability Data

Thus for a symmetric two-sided interval which contains at most probability 0.95, we take the interval

$$\{25, 26, \ldots, 54, 55\}$$

and hence define the rejection region

$$\mathcal{R} = \{16, 17, \ldots, 23, 24, 56, 57, \ldots, 64, 65\}$$

Note that this choice of rejection region ensures that there is at least probability 0.025 in each tail.

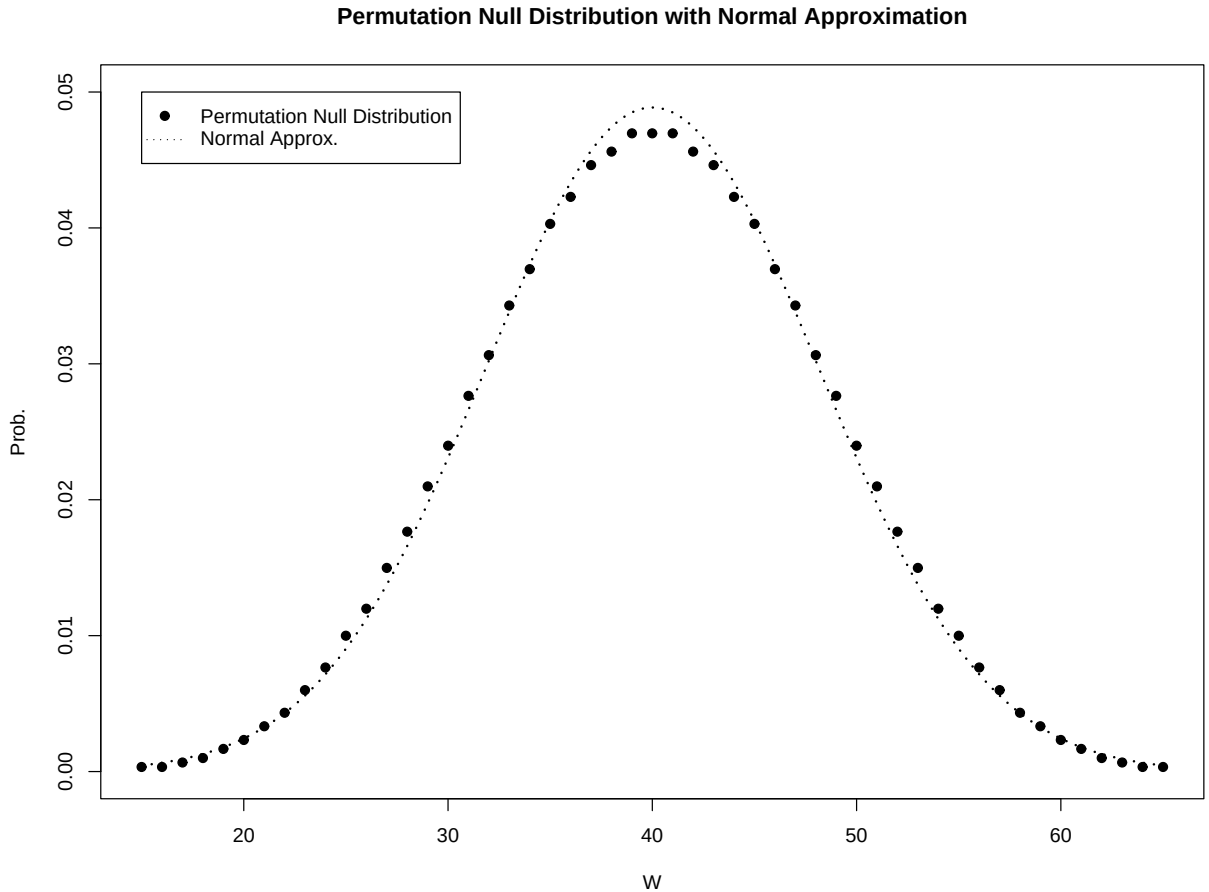
Thus in this example we do not reject the hypothesis of equal medians.

Normal Approximation

The Normal approximation to the exact null distribution is given by

$$W \dot{\sim} \text{Normal}\left(\frac{n_2(n_1+n_2+1)}{2}, \frac{n_1 n_2(n_1+n_2+1)}{12}\right)$$

Normal Approximation



Normal Approximation

The Normal approximation to the null distribution is quite good.

In fact, for most of the non-parametric tests, in large samples the Normal approximation yields a good approximation to the null distribution.

- Note that this Normal approximation is not the same as the Normality assumption we made in ANOVA and Regression settings
- The assumption for ANOVA and Regression related to each data point; the Normal approximation here relates to the behaviour of the test statistic.