Chapter 3 Completely Randomized Designs

Yibi Huang

3.1-3.8 Completely Randomized Designs

3.9 Experiments with Quantitative Factors, Goodness of Fit (Dose Response Modeling)

- Effects Model for CRD

Definition of a Completely Randomized Design (CRD) (1)

An experiment has a **completely randomized design** if

- the number of treatments g (including the *control* if there is one) is predetermined
- the number of replicates (n_i) in the ith treatment group is *predetermined*, $i = 1, \ldots, g$, and
- each allocation of $N = n_1 + \cdots + n_g$ experimental units into g groups of size $(n_1, \ldots, n_g)$ occurs equally likely.
- Say we have 4 units: A, B, C, D and, 2 treatments w/ 2 units each. The CRD ensures the following allocations occur equally likely

$$(AB, CD),\ (AC, BD),\ (AD, BC),$$
$$(BC, AD),\ (BD, AC),\ (CD, AB).$$

Definition of a Completely Randomized Design (CRD) (2)

- Tossing a coin for each of the 20 patients,

  if head $\longrightarrow$ treatment, if tail $\longrightarrow$ control

Definition of a Completely Randomized Design (CRD) (2)

- Tossing a coin for each of the 20 patients,

  $$\text{if head} \longrightarrow \text{treatment}, \quad \text{if tail} \longrightarrow \text{control}$$

  - NOT a CRD, as the number of replications in the 2 groups is not fixed.

Definition of a Completely Randomized Design (CRD) (2)

- Tossing a coin for each of the 20 patients,

  $$\text{if head} \longrightarrow \text{treatment}, \quad \text{if tail} \longrightarrow \text{control}$$

  - NOT a CRD, as the number of replications in the 2 groups is not fixed.

- If the patients draw lots, say, from 20 tickets in a hat, 10 of which are marked "treatment", it is a CRD.

Definition of a Completely Randomized Design (CRD) (2)

- Tossing a coin for each of the 20 patients,

  $$\text{if head} \longrightarrow \text{treatment}, \quad \text{if tail} \longrightarrow \text{control}$$

  - NOT a CRD, as the number of replications in the 2 groups is not fixed.

- If the patients draw lots, say, from 20 tickets in a hat, 10 of which are marked "treatment", it is a CRD.

- Tossing a coin for each of the 20 patients to determine who go to the treatment or the control group until one of the 2 groups reaches size 10, and all the unallocated patients go to the other group, it is a CRD.

Definition of a Completely Randomized Design (CRD) (2)

- Tossing a coin for each of the 20 patients,

  $$\text{if head} \longrightarrow \text{treatment}, \quad \text{if tail} \longrightarrow \text{control}$$

  - NOT a CRD, as the number of replications in the 2 groups is not fixed.

- If the patients draw lots, say, from 20 tickets in a hat, 10 of which are marked "treatment", it is a CRD.

- Tossing a coin for each of the 20 patients to determine who go to the treatment or the control group until one of the 2 groups reaches size 10, and all the unallocated patients go to the other group, it is a CRD.

- Say 10 of the patients are men and 10 are women. Randomly assign 5 men and 5 women to the treatment, and the rest to the control

Definition of a Completely Randomized Design (CRD) (2)

- Tossing a coin for each of the 20 patients,

  $$\text{if head} \longrightarrow \text{treatment}, \quad \text{if tail} \longrightarrow \text{control}$$

  - NOT a CRD, as the number of replications in the 2 groups is not fixed.

- If the patients draw lots, say, from 20 tickets in a hat, 10 of which are marked "treatment", it is a CRD.

- Tossing a coin for each of the 20 patients to determine who go to the treatment or the control group until one of the 2 groups reaches size 10, and all the unallocated patients go to the other group, it is a CRD.

- Say 10 of the patients are men and 10 are women. Randomly assign 5 men and 5 women to the treatment, and the rest to the control
  - this is a block design (in Ch13), NOT a CRD.

Definition of a Completely Randomized Design (CRD) (2)

- Tossing a coin for each of the 20 patients,

  $$\text{if head} \longrightarrow \text{treatment}, \quad \text{if tail} \longrightarrow \text{control}$$

  - NOT a CRD, as the number of replications in the 2 groups is not fixed.

- If the patients draw lots, say, from 20 tickets in a hat, 10 of which are marked "treatment", it is a CRD.

- Tossing a coin for each of the 20 patients to determine who go to the treatment or the control group until one of the 2 groups reaches size 10, and all the unallocated patients go to the other group, it is a CRD.

- Say 10 of the patients are men and 10 are women. Randomly assign 5 men and 5 women to the treatment, and the rest to the control
  - this is a block design (in Ch13), NOT a CRD.
  - both groups will have 5 men and 5 women. Using a CRD, the number of men and women in the groups may not be even

Means Model For A CRD Experiment

Consider an experiment with g treatments, and n_i replicates (i.e., n_i experimental units) for treatment i, $i = 1, 2, \ldots, g$.

$$\begin{array}{ll} \text{Treatment 1}: & y_{11}, y_{12}, \ldots, y_{1n_1} \\ \text{Treatment 2}: & y_{21}, y_{22}, \ldots\ldots\ldots, y_{2n_2} \\ & \vdots \\ \text{Treatment g}: & y_{g1}, y_{g2}, \ldots\ldots, y_{gn_g} \end{array}$$

$$\begin{array}{ccccccc} j\text{th unit for} & & \text{treatment} & & \text{error} & & \\ \text{treatment } i & & \text{effect} & & \text{(or noise)} & & \\ \downarrow & & \downarrow & & \downarrow & & i = 1, 2, \ldots, g \\ y_{ij} & = & \mu_i & + & \varepsilon_{ij} & & j = 1, 2, \ldots, n_i \end{array}$$

- μ_i = mean response for the ith treatment
- The error terms ε_{ij} are assumed to be **independent** with mean 0 and **constant variance** σ^2.
  Sometimes we further assume that errors are normal.

Dot and Bar Notation

A dot ($\bullet$) in subscript means *summing* over that index, for example

$$y_{i\bullet} = \sum_j y_{ij}, \quad y_{\bullet j} = \sum_i y_{ij}, \qquad y_{\bullet\bullet} = \sum_{i=1}^{g} \sum_{j=1}^{n_i} y_{ij}$$

A bar over a variable, along with a dot ($\bullet$) in subscript means *averaging* over that index, for example

$$\overline{y}_{i\bullet} = \frac{1}{n_i} \sum_{j=1}^{n_i} y_{ij}, \quad \overline{y}_{\bullet\bullet} = \frac{1}{N} \sum_{i=1}^{g} \sum_{j=1}^{n_i} y_{ij}$$

Parameter Estimation for the Means Model

Recall the least square estimates $\widehat{\mu}_i$'s are the $\widehat{\mu}_i$'s that *minimize* the *sum of squares* of the observations y_{ij} to their hypothesized means μ_i based on the model,

$$SS = \sum_{j=1}^{n_1}(y_{1j} - \widehat{\mu}_1)^2 + \sum_{j=1}^{n_2}(y_{2j} - \widehat{\mu}_2)^2 + \cdots + \sum_{j=1}^{n_g}(y_{gj} - \widehat{\mu}_g)^2.$$

In order to minimize S, we could differentiate it with respect to each μ_i and set the derivative equal to zero.

$$\frac{\partial S}{\partial \widehat{\mu}_i} = -2\sum_{j=1}^{n_i}(y_{ij} - \widehat{\mu}_i) = -2n_i(\overline{y}_{i\bullet} - \widehat{\mu}_i) = 0$$

The **least square estimate** for μ_i is thus the **sample mean** of observations in the corresponding treatment group,

$$\widehat{\mu}_i = \overline{y}_{i\bullet} = \frac{1}{n_i}\sum_{j=1}^{n_i} y_{ij}.$$

Moreover the LS estimate $\overline{y}_{i\bullet}$ for μ_i is **unbiased**.

Sum of Squares (1)

The means model for a CRD also have an a sum of squares identity very similar to $SST = SSR + SSE$ for a regression model.

$$y_{ij} - \overline{y}_{\bullet\bullet} = (\overline{y}_{i\bullet} - \overline{y}_{\bullet\bullet}) + (y_{ij} - \overline{y}_{i\bullet})$$

Squaring up both sides we get

$$(y_{ij} - \overline{y}_{\bullet\bullet})^2 = (\overline{y}_{i\bullet} - \overline{y}_{\bullet\bullet})^2 + (y_{ij} - \overline{y}_{i\bullet})^2 + 2(\overline{y}_{i\bullet} - \overline{y}_{\bullet\bullet})(y_{ij} - \overline{y}_{i\bullet})$$

Summing over the indexes we get

$$\overbrace{\sum_{i=1}^{g}\sum_{j=1}^{n_i}(y_{ij} - \overline{y}_{\bullet\bullet})^2}^{SST} = \overbrace{\sum_{i=1}^{g}\sum_{j=1}^{n_i}(\overline{y}_{i\bullet} - \overline{y}_{\bullet\bullet})^2}^{SSR=SS_{Trt}} + \overbrace{\sum_{i=1}^{g}\sum_{j=1}^{n_i}(y_{ij} - \overline{y}_{i\bullet})^2}^{SSE} + \underbrace{2\sum_{i=1}^{g}\sum_{j=1}^{n_i}(\overline{y}_{i\bullet} - \overline{y}_{\bullet\bullet})(y_{ij} - \overline{y}_{i\bullet})}_{= 0, \text{ see next slide}}$$

In this context of experimental design, SSR is often addressed as SS_{Trt}, called the **treatment sum of squares**

Sum of Squares (2)

Observe that

$$\sum_{i=1}^{g}\sum_{j=1}^{n_i}(\overline{y}_{i\bullet}-\overline{y}_{\bullet\bullet})(y_{ij}-\overline{y}_{i\bullet}) = \sum_{i=1}^{g}(\overline{y}_{i\bullet}-\overline{y}_{\bullet\bullet})\underbrace{\sum_{j=1}^{n_i}(y_{ij}-\overline{y}_{i\bullet})}$$

and

$$\sum_{j=1}^{n_i}(y_{ij}-\overline{y}_{i\bullet}) = y_{i\bullet}-n_i\overline{y}_{i\bullet} = y_{i\bullet}-n_i(\frac{y_{i\bullet}}{n_i}) = 0$$

and hence

$$\sum_{i=1}^{g}\sum_{j=1}^{n_i}(\overline{y}_{i\bullet}-\overline{y}_{\bullet\bullet})(y_{ij}-\overline{y}_{i\bullet}) = 0.$$

$$\underbrace{\sum_{i=1}^{g}\sum_{j=1}^{n_i}(y_{ij}-\overline{y}_{\bullet\bullet})^2}_{SST} = \underbrace{\sum_{i=1}^{g}\sum_{j=1}^{n_i}(\overline{y}_{i\bullet}-\overline{y}_{\bullet\bullet})^2}_{=SS_{Trt}=SSB} + \underbrace{\sum_{i=1}^{g}\sum_{j=1}^{n_i}(y_{ij}-\overline{y}_{i\bullet})^2}_{=SSE=SSW}$$

- SST = **total sum of squares**
  - reflects total variability in the response for all the units
- SS_{Trt} = **treatment sum of squares**
  - reflects variability **between** treatments
  - also called **between sum of squares**, denoted as SSB
- SSE = **error sum of squares**
  - Observe that SSE $= \sum_{i=1}^{g}(n_i-1)s_i^2$, in which

$$s_i^2 = \frac{1}{n_i-1}\sum_{j=1}^{n_i}(y_{ij}-\overline{y}_{i\bullet})^2$$

    is the sample variance **within** treatment group i.
    So SSE reflects the variability **within** treatment groups.
  - also called **within sum of squares**, denoted as SSW

Degrees of Freedom

Under the means model $y_{ij} = \mu_i + \varepsilon_{ij}$, and ε_{ij}'s i.i.d. $\sim N(0, \sigma^2)$, it can be shown that

$$\frac{\text{SSE}}{\sigma^2} \sim \chi^2_{N-g}.$$

Further assuming that $\mu_1 = \cdots = \mu_g$, then

$$\frac{\text{SST}}{\sigma^2} \sim \chi^2_{N-1}, \quad \frac{\text{SS}_{Trt}}{\sigma^2} \sim \chi^2_{g-1}$$

and SS_{Trt} is independent of SSE.
Note the **degrees of freedom** of the 3 SS

$$dfT = N - 1, \quad df_{Trt} = g - 1, \quad dfE = N - g$$

break down just like $SST = SS_{Trt} + SSE$,

$$dfT = df_{Trt} + dfE$$

Fitted Values, Residuals, and Estimate for σ^2

- **fitted value** for y_{ij} is $\widehat{y}_{ij} = \widehat{\mu}_i = \overline{y}_{i\bullet}$

- **residual** for y_{ij} is $e_{ij} = y_{ij} - \widehat{y}_{ij} = y_{ij} - \overline{y}_{i\bullet}$

- The estimate for σ^2 is again the MSE

$$\widehat{\sigma}^2 = MSE = \frac{SSE}{dfE}$$
$$= \frac{1}{N-g}\sum_{i=1}^{g}\sum_{j=1}^{n_i} e_{ij}^2 = \frac{1}{N-g}\sum_{i=1}^{g}\sum_{j=1}^{n_i}(y_{ij} - \overline{y}_{i\bullet})^2.$$

ANOVA F-test and ANOVA Table

To test whether the treatments have different effects

$H_0 : \mu_1 = \cdots = \mu_g$ (no difference between treatments)

$H_a : \mu_i$'s not all equal (some difference between treatments)

the test statistic is the F-statistic.

$$F = \frac{MS_{Trt}}{\text{MSE}} = \frac{SS_{Trt}/(g-1)}{SSE/(N-g)}$$

which has an F distribution with $g-1$ and $N-g$ degrees of freedom.

Source	Sum of Squares	d.f.	Mean Squares	F_0
Treatments	SS_{Trt}	$g-1$	$\text{MS}_{Trt} = \frac{SS_{Trt}}{g-1}$	$\frac{MS_{Trt}}{\text{MSE}}$
Errors	SSE	$N-g$	$\text{MSE} = \frac{SSE}{N-g}$	
Total	SST	$N-1$		

Interpretation of the ANOVA F-Statistic

$H_0 : \mu_1 = \cdots = \mu_g$ (no difference between treatments)

$H_a : \mu_i$'s not all equal (some difference between treatments)

$$F = \frac{SS_{Trt}/(g-1)}{SSE/(N-g)} = \frac{SSB/(g-1)}{SSW/(N-g)}$$
$$= \frac{\text{Variation Between Groups}}{\text{Variation Within Groups}}$$

The larger the variation between groups relative to variation within each group, the stronger the evidence toward H_a

Example — Resin Glue Failure Time — Background

- How to measure the lifetime of things like computer disk drives, light bulbs, and glue bonds?
  E.g., a computer drive is claimed to have a lifetime of 800,000 hours ($>$ 90 years).
  Clearly the manufacturer did not have disks on test for 90 years; how do they make such claims?

- *Accelerated life test*: Parts under stress (higher load, higher temperature, etc.) will usually fail sooner than parts that are unstressed. By modeling the lifetimes of parts under various stresses, we can estimate (extrapolate to) the lifetime of parts that are unstressed.

- Example: resin glue failure time

Example — Resin Glue Failure Time[1]

- Goal: to estimate the life time (in hours) of an encapsulating resin for gold-aluminum bonds in integrated circuits (operating at 120°C)
- Method: accelerated life test
- Design: Randomly assign 37 units to one of 5 different temperature stresses (in Celsius)

$$175^\circ, 194^\circ, 213^\circ, 231^\circ, 250^\circ$$

- Treatments: temperature in Celsius
- Response: $Y = \log_{10}$(time to failure in hours) of the tested material.

[1] Source: p. 448-449, *Accelerated Testing* (Nelson 2004). Original data is provided by Dr. Muhib Khan of AMD.

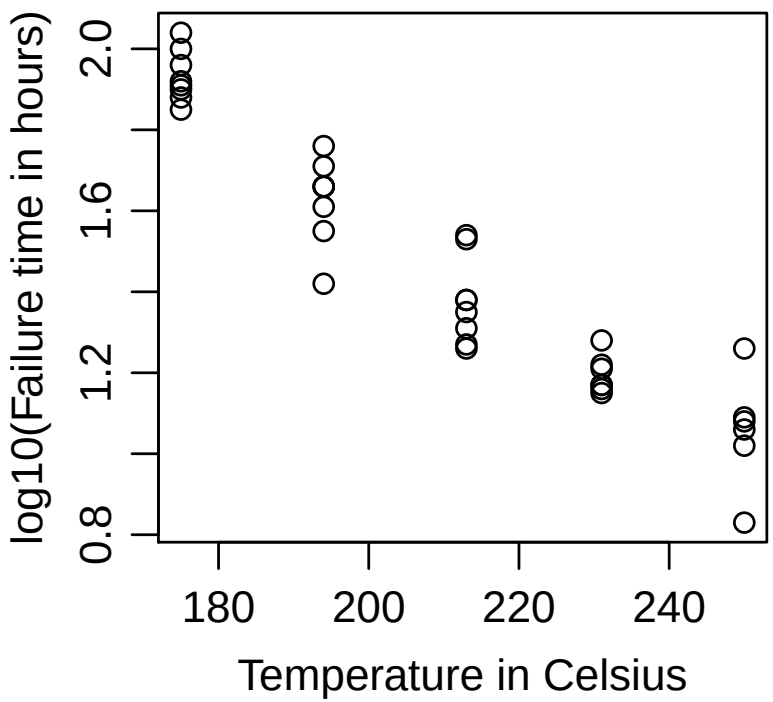
Example — Resin Glue Failure Time — Data

Temperature (°C)	175	194	213	231	250
Y	2.04	1.66	1.53	1.15	1.26
	1.91	1.71	1.54	1.22	0.83
	2.00	1.42	1.38	1.17	1.08
	1.92	1.76	1.31	1.16	1.02
	1.85	1.66	1.35	1.21	1.09
	1.96	1.61	1.27	1.28	1.06
	1.88	1.55	1.26	1.17	
	1.90	1.66	1.38		

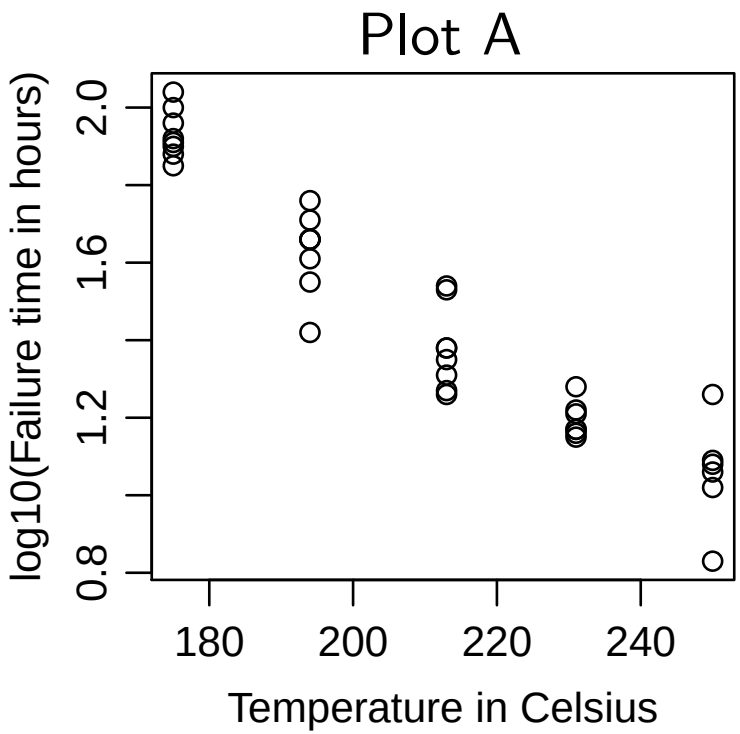
Data file: `resin2017.txt`

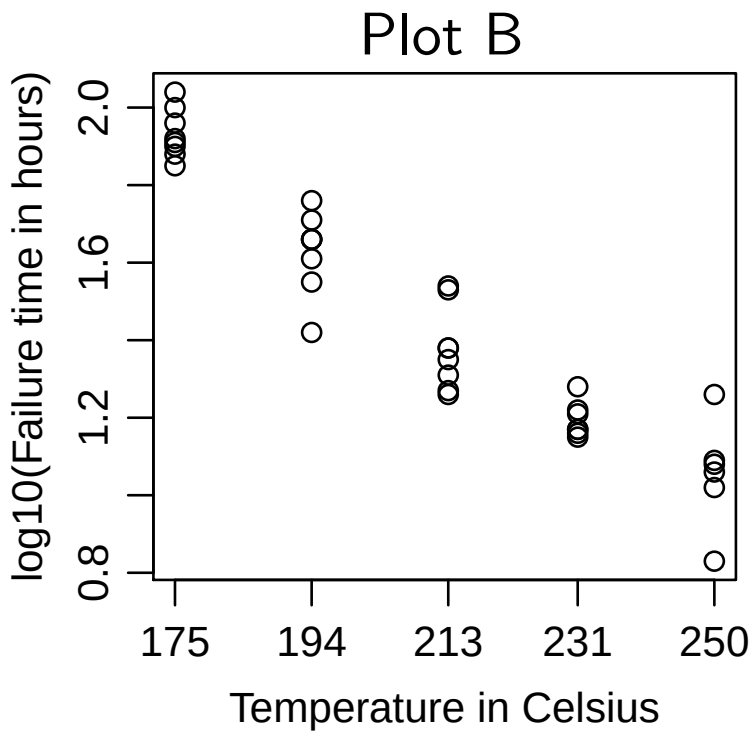
A First Look at the Data

```
> resin = read.table("resin.txt", header=T)
> str(resin)
'data.frame': 37 obs. of  2 variables:
 $ tempC: int  175 175 175 175 175 175 175 175 194 194 ...
 $ y    : num  2.04 1.91 2 1.92 1.85 1.96 1.88 1.9 1.66 1.71 ...

> plot(resin$tempC,resin$y, ylab="log10(Failure time in hours)",
  xlab="Temperature in Celsius")
```



```
plot(tempC,y,ylab="log10(Failure time in hours)",
 xlab="Temperature in Celsius")  # Plot A

plot(tempC,y,ylab="log10(Failure time in hours)",
 xlab="Temperature in Celsius",xaxt="n")
axis(side=1,at=c(175, 194, 213, 231, 250)) # Plot B
```





Which plot conveys more information?

Example — Resin Glue Failure Time — SS_{trt}

Temperature (°C)	175	194	213	231	250
y_{ij}	2.04	1.66	1.53	1.15	1.26
	1.91	1.71	1.54	1.22	0.83
	2.00	1.42	1.38	1.17	1.08
	1.92	1.76	1.31	1.16	1.02
	1.85	1.66	1.35	1.21	1.09
	1.96	1.61	1.27	1.28	1.06
	1.88	1.55	1.26	1.17	
	1.90	1.66	1.38		
n_i	8	8	8	7	6
$\overline{y}_{i\bullet}$	1.933	1.629	1.378	1.194	1.057

$$\overline{y}_{\bullet\bullet} = \tfrac{1}{37}(2.04 + 1.91 + \cdots + 1.06) = 1.465$$

The **between** group sum of squares

$$SS_{Trt} = \sum_{i=1}^{g}\sum_{j=1}^{n_i}(\overline{y}_{i\bullet} - \overline{y}_{\bullet\bullet})^2 = \sum_{i=1}^{5} n_i(\overline{y}_{i\bullet} - \overline{y}_{\bullet\bullet})^2$$

$$= 8(1.933 - 1.465)^2 + 8(1.629 - 1.465)^2 + 8(1.378 - 1.465)^2$$

$$+ 7(1.194 - 1.465)^2 + 6(1.057 - 1.465)^2 \approx 3.54$$

Example — Resin Glue Failure Time — SSE

The within group sum of squares:

$$\sum_{j=1}^{n_1}(y_{1j}-\overline{y}_{1\bullet})^2 = (2.04-1.933)^2+(1.91-1.933)^2+\cdots+(1.90-1.933)^2$$

$$\sum_{j=1}^{n_2}(y_{2j}-\overline{y}_{2\bullet})^2 = (1.66-1.629)^2+(1.71-1.629)^2+\cdots+(1.66-1.629)^2$$

$$\vdots \quad \vdots$$

$$\sum_{j=1}^{n_5}(y_{5j}-\overline{y}_{5\bullet})^2 = (1.26-1.057)^2+(0.83-1.057)^2+\cdots+(1.06-1.057)^2$$

So

$$SSE = \sum_{j=1}^{n_1}(y_{1j}-\overline{y}_{1\bullet})^2+\sum_{j=1}^{n_2}(y_{2j}-\overline{y}_{2\bullet})^2+\cdots+\sum_{j=1}^{n_5}(y_{5j}-\overline{y}_{5\bullet})^2 \approx 0.294$$

Example — Resin Glue Failure Time — F-statistic

- The observed F-statistic is

$$F_0 = \frac{SS_{Trt}/(g-1)}{SSE/(N-g)} = \frac{3.54/(5-1)}{0.29/(37-5)} \approx 97.66$$

- The resulting F-statistic is very large compared to 1. In fact, the p-value

$$P(F_{4,32} \geq 97.66) = 1.842 \times 10^{-17} \ll 0.001$$

- The data exhibit strong evidence against the null hypothesis that all means are equal.

ANOVA F-Test in R

```
> lm1 = lm(y ~ as.factor(tempC), data=resin)
> anova(lm1)
Analysis of Variance Table

Response: y
                 Df Sum Sq Mean Sq F value    Pr(>F)
as.factor(tempC)  4 3.5376 0.88441  96.363 < 2.2e-16 ***
Residuals        32 0.2937 0.00918
```

Always Check Degrees of Freedom

Without `as.factor()`, R will treat `tempC` as numeric variable and fit the model

$$y_{ij} = \beta_0 + \beta_1 \texttt{tempC}_i + \varepsilon_{ij}.$$

rather than the means model $y_{ij} = \mu_i + \varepsilon_{ij}$.

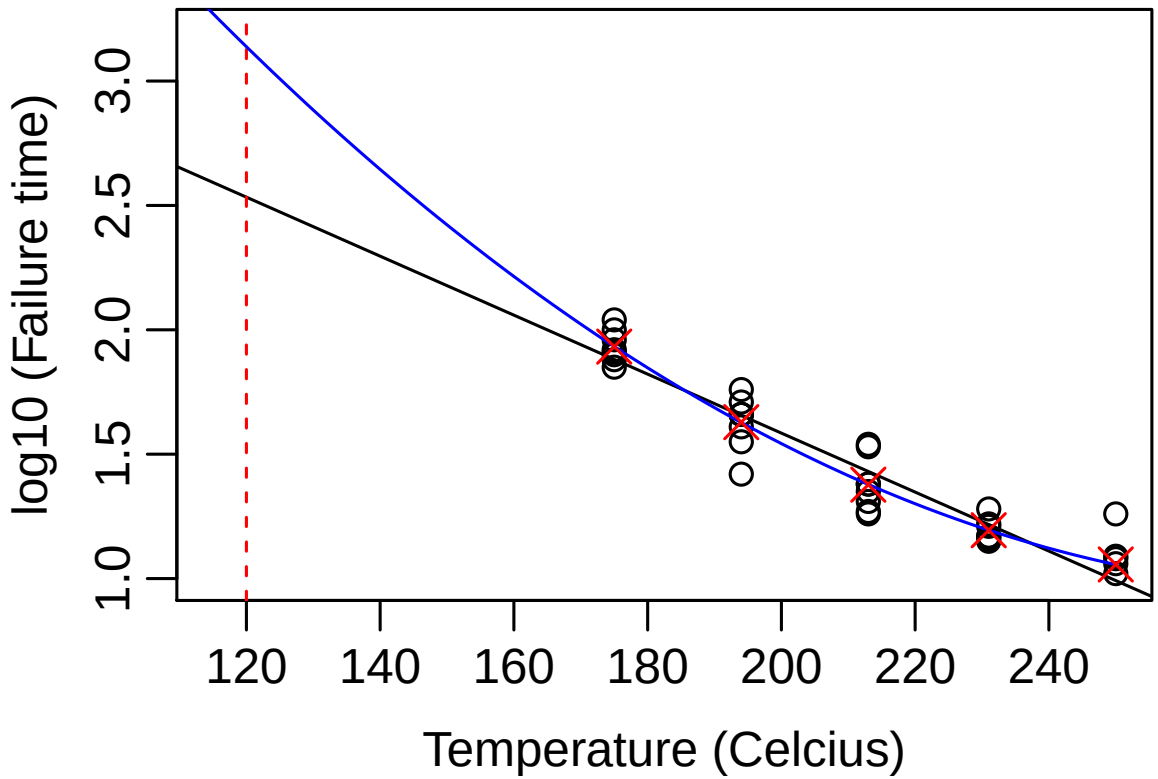
```
> lm2 = lm(y ~ tempC, data=resin)
> anova(lm2)
Analysis of Variance Table

Response: y
          Df Sum Sq Mean Sq F value    Pr(>F)
tempC      1 3.4593  3.4593  325.41 < 2.2e-16 ***
Residuals 35 0.3721  0.0106
```

The easiest way to know if R fits the right model is to check the degrees of freedom in the output. In the table above, the d.f. for temp should be $5 - 1 = 4$ rather than 1.

Limitation of ANOVA F-Tests

The ANOVA F-test merely tells us the glue has different failure time at different temperature.

However, our goal is to predict the lifetime of the glue at a temperature of 120°.



(red cross marks group means)

R Codes for Plot on the Previous Page

```
lm1 = lm(y ~ tempC, data=resin)
lm2 = lm(y ~ tempC + I(tempC^2), data=resin)
plot(resin$tempC, resin$y, ylab="log10 (Failure time)",xlab="Temperatur
     xlim=c(115,250),ylim=c(1,3.2))

# Adding the vertical dash line at tempC=120
abline(v=120, col=2, lty=2)

# Adding the linear regression line
abline(lm1)

# Adding the fitted curve for the quadratic function
curve(predict(lm2, newdata=data.frame(tempC=x)), col=4, add=T)

# Marking the group means by a red cross
points(c(175, 194, 213, 231, 250), tapply(y, tempC, mean),
       col=2, pch=4, cex=1.5)
```

Dose-Response Modeling

In some experiments, the treatments are associated with *numerical levels* x_i such as drug dose, baking time, or temperature.

We will refer to such levels as *doses*.

- The means model $y_{ij} = \mu_i + \varepsilon_{ij}$ specifies **no relationship** between treatment levels x_i and the response y, which cannot be used to infer the response at some dose x other than those used in the experiment
- With a *quantitative* treatment factor, experimenters are usually more interested on how the response is affected by the factor as a function of x_i

$$y_{ij} = f(x_i; \theta) + \varepsilon_{ij},$$

e.g.,

$$f(x_i; \beta_0, \beta_1) = \beta_0 + \beta_1 x_i;$$
$$f(x_i; \beta_0, \beta_1, \beta_2) = \beta_0 + \beta_1 x_i + \beta_2 x_i^2; \text{ or}$$
$$f(x_i; \beta_0, \beta_1) = \beta_0 + \beta_1 \log(x_i).$$

$$y_{ij} = f(x_i; \theta) + \varepsilon_{ij}$$

Advantages of dose-response modeling

- less complex (fewer parameters)
- easier to interpret (sometimes)
- generalizable to doses not included in the experiment

Issues to consider:

- How to choose the function f?
  - One commonly used family of functions f are *polynomials*:

$$f(x_i; \beta) = \beta_0 + \beta_1 x_i + \beta_2 x_i^2 + \cdots + \beta_k x_i^k,$$

    But polynomials are NOT always the best choice
  - For simplicity, we would choose the lowest possible order of polynomial that adequately fit the data.
- How to assess how well f fits the data? . . . . . Goodness of fit

Polynomial Models

Let t_i denote the temperature in Celsius in treatment group i. Consider the following polynomial models for the resin glue data.

$$\begin{aligned}
\text{Null Model} &: y_{ij} = \mu + \varepsilon_{ij} \\
\text{Linear Model} &: y_{ij} = \beta_0 + \beta_1 t_i + \varepsilon_{ij} \\
\text{Quadratic Model} &: y_{ij} = \beta_0 + \beta_1 t_i + \beta_2 t_i^2 + \varepsilon_{ij} \\
\text{Cubic Model} &: y_{ij} = \beta_0 + \beta_1 t_i + \beta_2 t_i^2 + \beta_3 t_i^3 + \varepsilon_{ij} \\
\text{Quartic Model} &: y_{ij} = \beta_0 + \beta_1 t_i + \beta_2 t_i^2 + \beta_3 t_i^3 + \beta_4 t_i^4 + \varepsilon_{ij}
\end{aligned}$$

- Every model is nested in the model below it. (Why?)
- Don't skip terms. If a higher order term is significant, e.g., t_i^3, than all lower order terms have to be kept (1, t_i, t_i^2), even if they are not significant.
- Why no quintic or higher order models?

In general, for an experiment with g treatment groups, if the treatment factor is numeric, one can fit a polynomial model up to degree $g-1$

$$y_{ij} = \beta_0 + \beta_1 x_i + \cdots + \beta_{g-1} x_i^{g-1} + \varepsilon_{ij}.$$

Question: For the resin glue data, what will happen if a quintic model (a polynomial of order 5) is fitted?

$$y_{ij} = \beta_0 + \beta_1 t_i + \beta_2 t_i^2 + \beta_3 t_i^3 + \beta_4 t_i^4 + \beta_5 t_i^5 + \varepsilon_{ij}$$

In general, for an experiment with g treatment groups, if the treatment factor is numeric, one can fit a polynomial model up to degree $g-1$

$$y_{ij} = \beta_0 + \beta_1 x_i + \cdots + \beta_{g-1} x_i^{g-1} + \varepsilon_{ij}.$$

Question: For the resin glue data, what will happen if a quintic model (a polynomial of order 5) is fitted?

$$y_{ij} = \beta_0 + \beta_1 t_i + \beta_2 t_i^2 + \beta_3 t_i^3 + \beta_4 t_i^4 + \beta_5 t_i^5 + \varepsilon_{ij}$$

Answer: There exist more than one polynomial of degree 5 passing through the 5 points $(175, \mu_1)$, $(194, \mu_2)$, $(213, \mu_3)$, $(231, \mu_4)$, and $(250, \mu_5)$. Thus the 6 coefficients $\beta_0, \beta_1, \ldots, \beta_5$ CANNOT be uniquely determined.

As a rule of thumb, for an experiment with g treatments, we can fit a model with at most g parameters.

Linear Model (1)

Let's try to fit the linear model: $y_{ij} = \beta_0 + \beta_1 t_i + \varepsilon_{ij}$.

```
> lm1 = lm(y ~ tempC, data = resin)
> summary(lm1)
             Estimate Std. Error t value Pr(>|t|)
(Intercept)  3.9560075  0.1391174   28.44   <2e-16 ***
tempC       -0.0118567  0.0006573  -18.04   <2e-16 ***
---
Residual standard error: 0.1031 on 35 degrees of freedom
Multiple R-squared: 0.9029,     Adjusted R-squared: 0.9001
F-statistic: 325.4 on 1 and 35 DF,  p-value: < 2.2e-16
```

- Fitted equation: $\log_{10}(\text{failure time}) = 3.956 - 0.01186\,T$
- Predicted $\log_{10}(\text{failure time})$ at 120° is

$$3.956 - 0.01186 \times 120 \approx 2.5332,$$

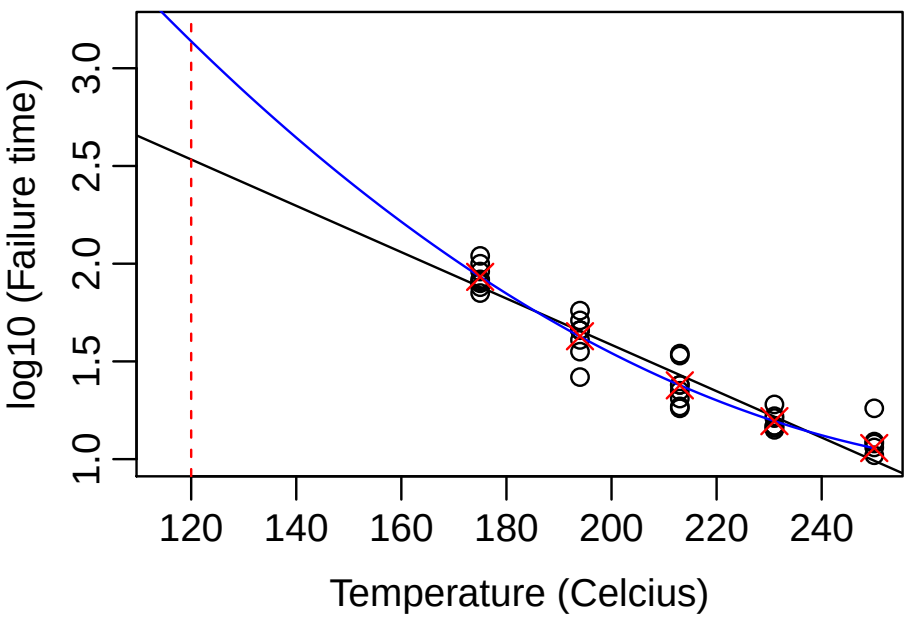
and hence the failure time at 120° is predicted as

$$10^{2.5332} \approx 341 \text{ hours.}$$

Linear Model (2)

R commands for the predicted log10(failure time) along with a 95% prediction interval:

```
> predict(lm1, newdata=data.frame(tempC=120), interval="prediction")
       fit      lwr      upr
1 2.533201 2.289392 2.777011
```



By imposing the regression line on the top of the scatter plot, we can see y is a slightly curved with temperature. Using the linear model, the failure time at 120° will be *underestimated*.

Quadratic Model

```
> lm2 = lm(y ~ tempC+I(tempC^2), data=resin)
> summary(lm2)
(... part of the output is omitted ...)
Coefficients:
               Estimate Std. Error t value Pr(>|t|)
(Intercept)   7.4179987  1.1564331   6.415 2.51e-07 ***
tempC        -0.0450981  0.0110542  -4.080 0.000258 ***
I((tempC)^2)  0.0000786  0.0000261   3.011 0.004879 **
---
Residual standard error: 0.09295 on 34 degrees of freedom
Multiple R-squared: 0.9233,     Adjusted R-squared: 0.9188
F-statistic: 204.8 on 2 and 34 DF,  p-value: < 2.2e-16
```

- Fitted model: $\log_{10}(\text{time}) = 7.418 - 0.0451\,T + 0.0000786\,T^2$
- Predicted log10(time) at 120° is

$$7.418 - 0.0451 \times 120 + 0.0000786 \times (120)^2 \approx 3.138$$

The predicted failure time at 120° is $10^{3.138} \approx 1374$ hours.

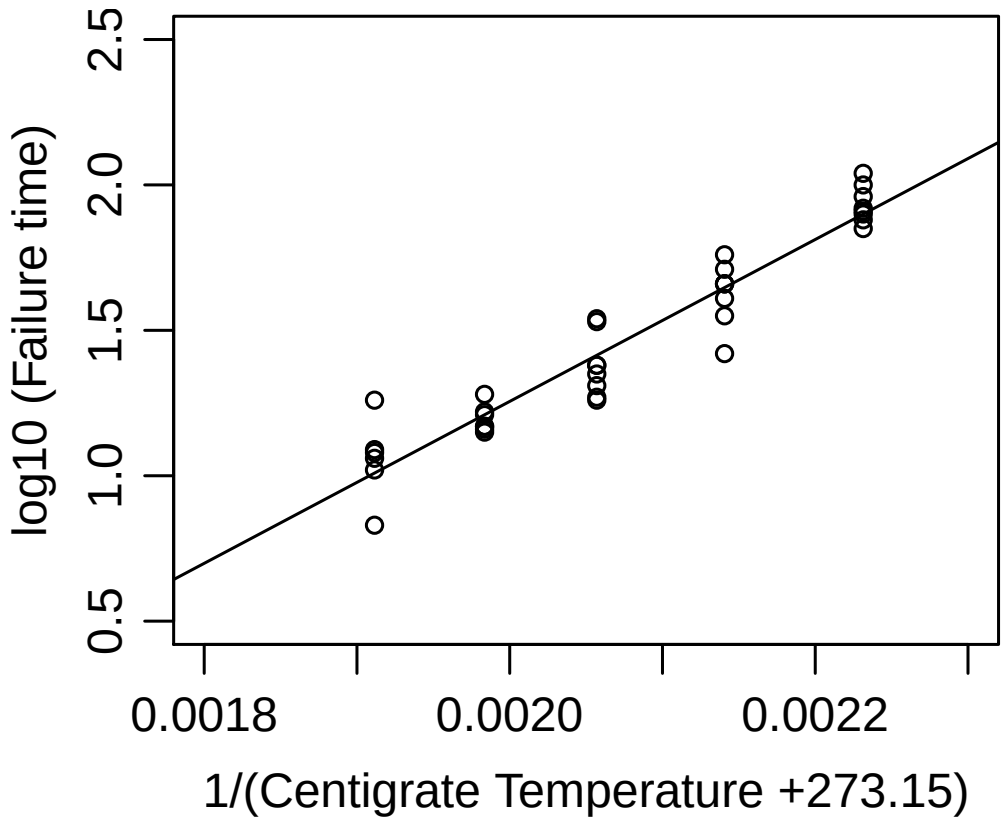
Cubic & Quartic Model

```
> lm3 = lm(y ~ tempC+I(tempC^2)+I(tempC^3), data = resin)
> summary(lm3)
             Estimate Std. Error t value Pr(>|t|)
(Intercept)  6.827e+00  1.299e+01   0.526    0.603
tempC       -3.659e-02  1.865e-01  -0.196    0.846
I(tempC^2)   3.815e-05  8.860e-04   0.043    0.966
I(tempC^3)   6.357e-08  1.392e-06   0.046    0.964

Residual standard error: 0.09434 on 33 degrees of freedom
Multiple R-squared:  0.9233,    Adjusted R-squared:  0.9164
F-statistic: 132.5 on 3 and 33 DF,  p-value: < 2.2e-16

> lm4 = lm(y ~ tempC+I(tempC^2)+I(tempC^3)+I(tempC^4), data = resin)
Coefficients:
             Estimate Std. Error t value Pr(>|t|)
(Intercept)  9.699e-01  1.957e+02   0.005    0.996
tempC        7.573e-02  3.750e+00   0.020    0.984
I(tempC^2)  -7.649e-04  2.679e-02  -0.029    0.977
I(tempC^3)   2.600e-06  8.459e-05   0.031    0.976
I(tempC^4)  -2.988e-09  9.962e-08  -0.030    0.976

Residual standard error: 0.0958 on 32 degrees of freedom
Multiple R-squared:  0.9233,    Adjusted R-squared:  0.9138
F-statistic: 96.36 on 4 and 32 DF,  p-value: < 2.2e-16
```

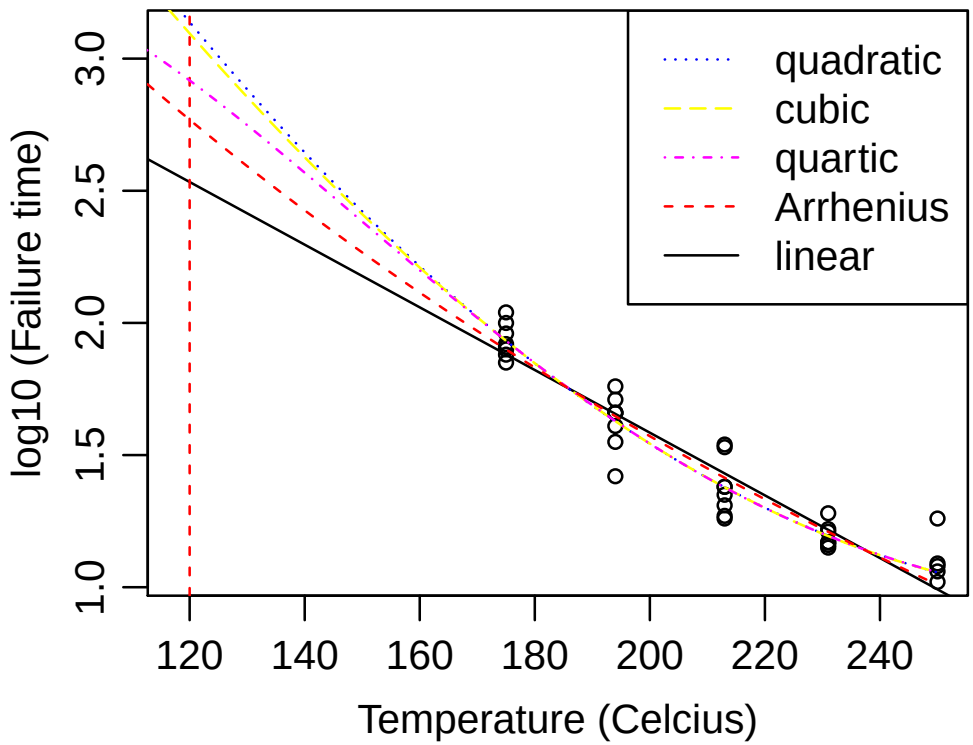
Arrhenius Law

The Arrhenius rate law in Thermodynamics says, the log of failure time is linear in the inverse of absolute Kelvin temperature, which equals the Centigrade temperature plus 273.16 degrees.

Arrhenius Model: $y_{ij} = \beta_0 + \dfrac{\beta_1}{T + 273.15}$.



```
> lmarr = lm(y ~ I(1/(tempC+273.15)), data=resin)
> summary(lmarr)
(... some output is omitted ...)
Coefficients:
                            Estimate Std. Error t value Pr(>|t|)
(Intercept)                  -4.3120     0.3007  -14.34  3.2e-16 ***
I(1/(tempC + 273.15))      2783.7764   144.6808   19.24  < 2e-16 ***

Residual standard error: 0.09724 on 35 degrees of freedom
Multiple R-squared: 0.9136,     Adjusted R-squared: 0.9112
F-statistic: 370.2 on 1 and 35 DF,  p-value: < 2.2e-16
```



Data Can Distinguish Models Only at Design Points

In addition to polynomial models and the Arrhenius model, many other models can be considered

$$
\begin{aligned}
y_{ij} &= \beta_0 + \beta_1 \log(t_i) + \varepsilon_{ij}, \\
y_{ij} &= \beta_0 + \beta_1 \exp(t_i) + \varepsilon_{ij}, \\
y_{ij} &= \beta_0 + \beta_1 \sin(t_i) + \varepsilon_{ij}, \\
y_{ij} &= \beta_0 + f(t_i) + \varepsilon_{ij}.
\end{aligned}
$$

As we only have observations at five temperatures:

$$175,\ 194,\ 213,\ 231,\ 250,$$

the data cannot distinguish between two models:

$$y_{ij} = f(t_i) + \varepsilon_{ij} \quad \text{and} \quad y_{ij} = g(t_i) + \varepsilon_{ij},$$

if $f(t)$ and $g(t)$ coincide at $t = 175, 194, 213, 231, 250$, even if f and g behave differently in other places.

The Model that Fit the Data the Best

If no restriction is placed on f, how well the model $y_{ij} = f(t_i) + \varepsilon_{ij}$ can possibly fit the data?

The least square method will choose the f that minimize

$$\sum_i \sum_j (y_{ij} - f(t_i))^2$$

Recall that given a list of numbers $x_1, x_2, \ldots, x_n$ the c that minimize $\sum_{i=1}^n (x_i - c)^2$ is the mean $\overline{x} = \frac{1}{n}\sum_{i=1}^n x_i$.

Thus the least square method will choose the f that

$$f(t_i) = \overline{y}_{i\bullet}.$$

Thus the smallest SSE a model $y_{ij} = f(t_i) + \varepsilon_{ij}$ can possibly achieve is

$$\sum_i \sum_j (y_{ij} - \overline{y}_{i\bullet})^2$$

which is the SSE for the **means model** $y_{ij} = \mu_i + \varepsilon_{ij}$.

Conclusion: **no other models can beat the means model in minimizing the SSE**.

Goodness of Fit

As the means model is the model that fit the data the best, we can access the goodness of a model $y_{ij} = f(t_i) + \varepsilon_{ij}$ by comparing it with the means model.

$$\text{Full Model} : y_{ij} = \mu_i + \varepsilon_{ij}$$
$$\text{Reduced Model} : y_{ij} = f(t_i) + \varepsilon_{ij}$$

This comparison is legitimate because any model $y_{ij} = f(t_i) + \varepsilon_{ij}$ is nested in the means model $y_{ij} = \mu_i + \varepsilon_{ij}$ (letting $\mu_i = f(t_i)$).

We can use the F-statistic below to compare a reduced model with a full model

$$F = \frac{(SSE_{reduced} - SSE_{full})/(df_{reduced} - df_{full})}{SSE_{full}/df_{full}}$$

Goodness of Fit of the Linear Model

Since the linear model (reduced model) is nested in the means model (full), use the F-statistic for model comparison we get

```
> lm1 = lm(y ~ tempC, data = resin)                 # linear model
> lmmeans = lm(y ~ as.factor(tempC), data = resin) # means model
> anova(lm1,lmmeans)
Analysis of Variance Table

Model 1: y ~ tempC
Model 2: y ~ as.factor(tempC)
  Res.Df     RSS Df Sum of Sq      F  Pr(>F)
1     35 0.37206
2     32 0.29369  3   0.07837 2.8463 0.05303 .
```

The P-value 0.05303 is moderate evidence showing the linear doesn't fit the data so well.

Goodness of Fit of the Quadratic Model

Since the quadratic model (reduced model) is also nested in the means model (full model), again use the F-statistic for model comparison we get

```
> lm2 = lm(y ~ tempC+I((tempC)^2), data=resin)      # quadratic model
> lmmeans = lm(y ~ as.factor(tempC), data = resin)  # means model
> anova(lm2,lmmeans)
Analysis of Variance Table

Model 1: y ~ tempC + I((tempC)^2)
Model 2: y ~ as.factor(tempC)
  Res.Df     RSS Df  Sum of Sq      F Pr(>F)
1     34 0.29372
2     32 0.29369  2 2.6829e-05 0.0015 0.9985
```

The large p-value 0.9985 shows the quadratic model fits the data nearly as good as the best model. Thus, the quadratic seems to be an appropriate model for the data.

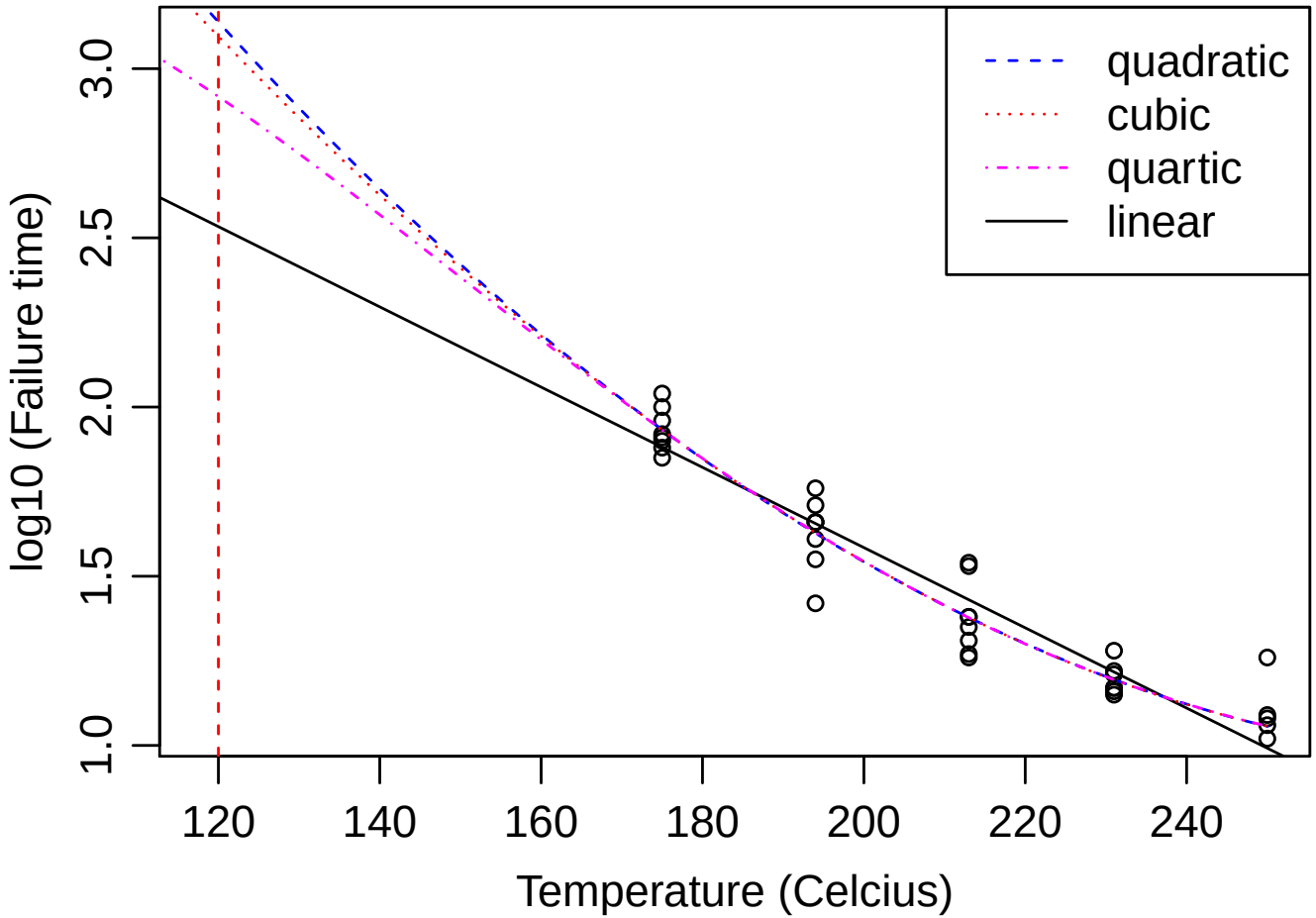
Shall We Consider a Cubic or a Quartic Model?

No. Because

$$\text{Quadratic} \subset \text{Cubic} \subset \text{Quartic} \subset \text{Means Model}$$

the cubic model and quartic model won't fit the data better than the means model does. As the quadratic model fits the data nearly as well as the means model, the 4 models just fit as well as each other. In this case we simply choose the model of lowest complexity.

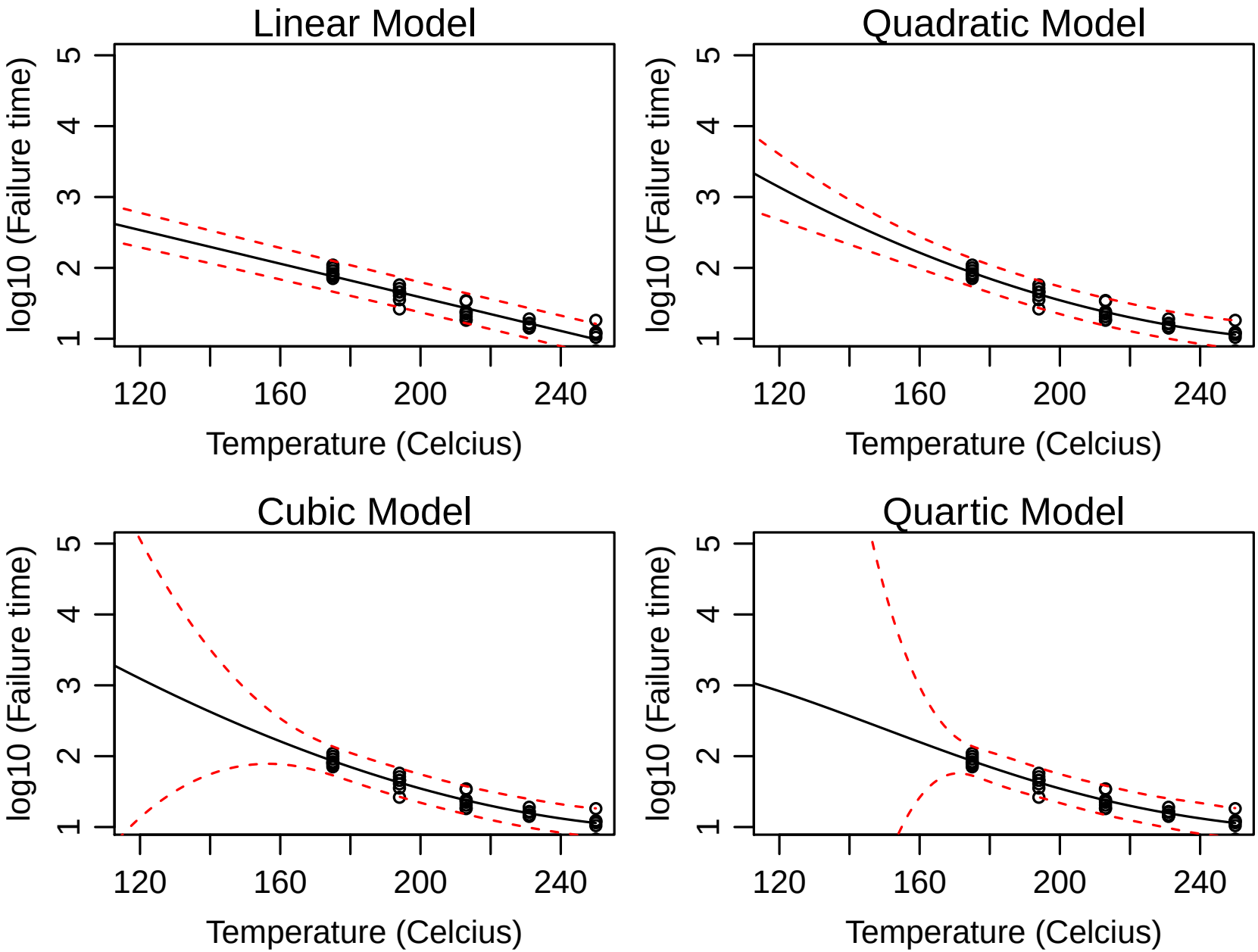
Be Cautious About Extrapolation



Though the quadratic, cubic, and quartic model fit the 5 points nearly as well, their predicted values at 120°C are quite different,

$$\text{quadratic} > \text{cubic} > \text{quartic} > \text{linear}$$

95% Prediction Intervals



Prediction Intervals at 120°C

Observe the length of the 95% prediction interval increase with the degree of the polynomial.

```
> predict(lm1, newdata=data.frame(tempC=120),interval="p")
       fit      lwr      upr
1 2.533201 2.289392 2.777011

> predict(lm2, newdata=data.frame(tempC=120),interval="p")
       fit      lwr      upr
1 3.138128 2.674383 3.601874

> predict(lm3, newdata=data.frame(tempC=120),interval="p")
       fit      lwr      upr
1 3.095342 1.132382 5.058303

> predict(lm4, newdata=data.frame(tempC=120),interval="p")
       fit       lwr      upr
1 2.917399 -9.330658 15.16546
```

Though the quadratic, cubic, and quartic models fit in the range of data points (175°C - 250°C) as well as each other, outside that range the reliability of their prediction changes drastically.

Since the Arrhenius model is nested in the means model, we can check its goodness of fit.

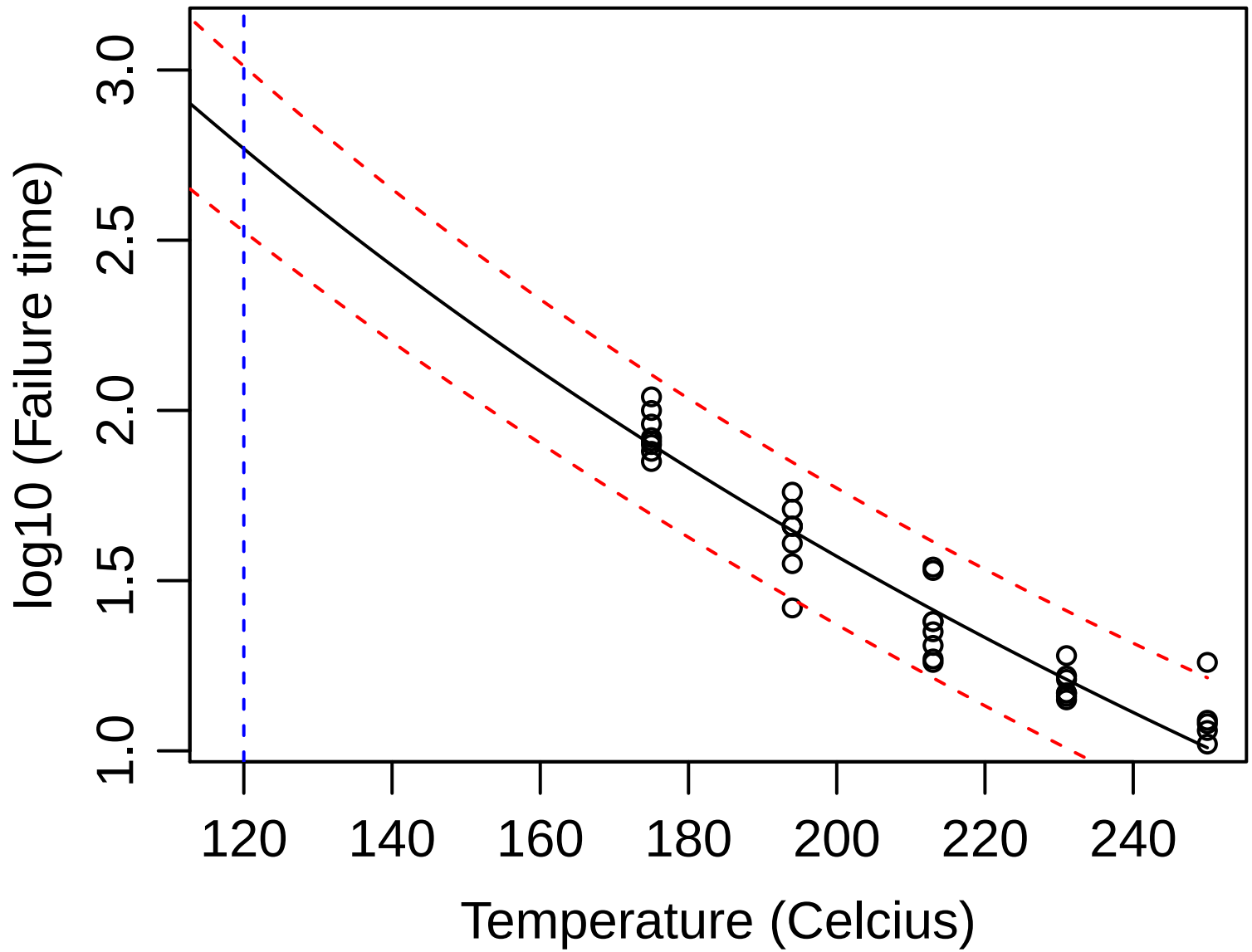
```
> anova(lmarr,lmmeans)
Analysis of Variance Table

Model 1: y ~ I(1/(tempC + 273.15))
Model 2: y ~ as.factor(tempC)
  Res.Df     RSS Df Sum of Sq      F Pr(>F)
1     35 0.33093
2     32 0.29369  3  0.037239 1.3525 0.2749
```

The moderately large P-value 0.2749 told us the Arrhenius Model is acceptable relative to the best model.

```
> predict(lmarr, newdata=data.frame(tempC=120),interval="p")
      fit      lwr      upr
1 2.76868 2.525909 3.011451
```

So the predicted failure time at 120°C is $10^{2.769} = 587.0571$ hours, and the 95% prediction interval is $(10^{2.526}, 10^{3.011}) \approx (336, 1026)$ hours.

95% Prediction Interval Based on the Arrhenius Model



Effects Model for CRD

Means Model Is a Multiple Linear Regression Model

For an experiment with g treatments, the Means model

$$y_{ij} = \mu_i + \varepsilon_{ij}$$

can be written as a multiple linear regression model by defining a *dummy variable* for each treatment group. The dummy variable for the ith treatment is defined as

$$D_i = \begin{cases} 1 & \text{if the experimental unit receives the } i\text{th treatment} \\ 0 & \text{otherwise} \end{cases}$$

The means model can then be written as a regression model

$$Y_k = \mu_1 D_{1k} + \mu_2 D_{2k} + \cdots + \mu_g D_{gk} + \varepsilon_k$$

Note that this regression model has *no intercept*.

In R, putting -1 in the model formula tells R to fit a regression model with no intercept.

```
> resin = read.table("resin2017.txt", header=T)
> lmmeans = lm(y ~ -1 + as.factor(tempC), data = resin)
> lmmeans

Call:
lm(formula = y ~ -1 + as.factor(tempC), data = resin)

Coefficients:
as.factor(tempC)175  as.factor(tempC)194  as.factor(tempC)213
              1.932                 1.629                 1.377
as.factor(tempC)231  as.factor(tempC)250
              1.194                 1.057
```

Recall for the resin glue data, the group means $\overline{y}_{i\bullet}$ are

Temperature (°C)	175	194	213	231	250
$\overline{y}_{i\bullet}$	1.933	1.629	1.378	1.194	1.057

Observed the coefficients 1.932, 1.629, ..., etc, are simply the group means $\overline{y}_{i\bullet}$.

The command `as.factor(tempC)` tells R to create *dummy variables* for each levels the temperature. Without `as.factor()`, R will fit the model

$$y_{ij} = \beta t_i + \varepsilon_{ij}$$

where t_i is the temperature in Celsius for treatment group i.

```
> lmmeans1 = lm(y ~ -1 + tempC, data = resin)
> lmmeans1

Call:
lm(formula = y ~ -1 + tempC, data = resin)

Coefficients:
   tempC
0.006695
```

Effects Model for a CRD Experiments

The textbook formulate the means model in another form:

$$\begin{aligned} y_{ij} &= \mu_i + \varepsilon_{ij} && \text{(means model)} \\ &= \mu + \alpha_i + \varepsilon_{ij} && \text{(effects model)} \end{aligned}$$

- Observe the effects model has $g + 1$ parameters $\mu, \alpha_1, \ldots, \alpha_g$, while the means model only has g parameters $\mu_1, \ldots, \mu_g$
- The effects model is **overparameterized**, meaning that it specifies more parameters than we actually need. The two sets of parameters below

$$(\mu, \alpha_1, \ldots, \alpha_g) \quad \text{and} \quad (\mu - c, \alpha_1 + c, \ldots, \alpha_g + c)$$

  specifies identical means for the responses. Thus the parameters for the effects model **cannot be uniquely determined**.
- These two models are equivalent in the sense that fitted values for responses will be identical.

How to Deal With Overparametrization?

There are various ways to deal with overparametrization in the effects model

$$y_{ij} = \mu + \alpha_i + \varepsilon_{ij}$$

Some common ways include

- letting $\mu = 0$

How to Deal With Overparametrization?

There are various ways to deal with overparametrization in the effects model

$$y_{ij} = \mu + \alpha_i + \varepsilon_{ij}$$

Some common ways include

- letting $\mu = 0$
- letting one of the α_i's to be 0

How to Deal With Overparametrization?

There are various ways to deal with overparametrization in the effects model

$$y_{ij} = \mu + \alpha_i + \varepsilon_{ij}$$

Some common ways include

- letting $\mu = 0$
- letting one of the α_i's to be 0
- letting $\alpha_1 + \alpha_2 + \ldots + \alpha_g = 0$

How to Deal With Overparametrization?

There are various ways to deal with overparametrization in the effects model

$$y_{ij} = \mu + \alpha_i + \varepsilon_{ij}$$

Some common ways include

- letting $\mu = 0$
- letting one of the α_i's to be 0
- letting $\alpha_1 + \alpha_2 + \ldots + \alpha_g = 0$
  - lack natural interpretation

How to Deal With Overparametrization?

There are various ways to deal with overparametrization in the effects model

$$y_{ij} = \mu + \alpha_i + \varepsilon_{ij}$$

Some common ways include

- letting $\mu = 0$
- letting one of the α_i's to be 0
- letting $\alpha_1 + \alpha_2 + \ldots + \alpha_g = 0$
  - lack natural interpretation
  - give simpler formulas when CRD is generalized to **factorial designs** (Chapter 8).

How to Deal With Overparametrization?

There are various ways to deal with overparametrization in the effects model

$$y_{ij} = \mu + \alpha_i + \varepsilon_{ij}$$

Some common ways include

- letting $\mu = 0$
- letting one of the α_i's to be 0
- letting $\alpha_1 + \alpha_2 + \ldots + \alpha_g = 0$
  - lack natural interpretation
  - give simpler formulas when CRD is generalized to **factorial designs** (Chapter 8).
- letting $n_1\alpha_1 + n_2\alpha_2 + \ldots + n_g\alpha_g = 0$
  Here n_i is the number of replicates in the ith treatment group

How to Deal With Overparametrization?

There are various ways to deal with overparametrization in the effects model

$$y_{ij} = \mu + \alpha_i + \varepsilon_{ij}$$

Some common ways include

- letting $\mu = 0$
- letting one of the α_i's to be 0
- letting $\alpha_1 + \alpha_2 + \ldots + \alpha_g = 0$
  - lack natural interpretation
  - give simpler formulas when CRD is generalized to **factorial designs** (Chapter 8).
- letting $n_1\alpha_1 + n_2\alpha_2 + \ldots + n_g\alpha_g = 0$
  Here n_i is the number of replicates in the ith treatment group
  - Under this constraint, least square estimates for μ and α_i's have the simple form

$$\widehat{\mu} = \overline{y}_{\bullet\bullet}, \qquad \widehat{\alpha}_i = \overline{y}_{i\bullet} - \overline{y}_{\bullet\bullet}$$

When One of the α_i's is Dropped ...

Say α_1 is dropped, the mean response for the g treatments are

$$\mathbb{E}[y_{ij}] = \begin{cases} \mu & \text{for treatment 1} \\ \mu + \alpha_2 & \text{for treatment 2} \\ \vdots & \\ \mu + \alpha_g & \text{for treatment g} \end{cases}$$

- The mean response under the first treatment ($i = 1$) is μ
- $\alpha_i =$ the difference between the mean response of the ith treatment and that of the 1st treatment. One can compare the effect of the ith treatment and the 1st treatment by testing $\alpha_i = 0$
- Useful for comparing treatments

```
> lmeffects1 = lm(y ~ as.factor(tempC), data = resin)
> lmeffects1

Call:
lm(formula = y ~ as.factor(tempC), data = resin)

Coefficients:
        (Intercept)  as.factor(tempC)194  as.factor(tempC)213
             1.9325              -0.3037              -0.5550
as.factor(tempC)231  as.factor(tempC)250
            -0.7382              -0.8758
```

Note there is no `as.factor(temp)175`, $\widehat{\alpha}_1$, since R sets $\alpha_1 = 0$.

Temperature (°C)	175	194	213	231	250
$\overline{y}_{i\bullet}$	1.933	1.629	1.378	1.194	1.057

Observed $\widehat{\mu} = \overline{y}_{1\bullet}$ and $\widehat{\alpha}_i = \overline{y}_{i\bullet} - \overline{y}_{1\bullet}$.