Unbalanced designs and polynomial regression

Session 8

MATH 80667A: Experimental Design and Statistical Methods
for Quantitative Research in Management
HEC Montréal

Outline

Unbalanced designs

Polynomial regression

Unbalanced designs

Premise

So far, we have exclusively considered balanced samples

balanced = same number of observational units in each sub-groups

Most experiments (even planned) end up with unequal sample sizes.

Noninformative drop-out

Unbalanced samples may be due to many causes, including randomization (need not balance) and loss-to-follow up (dropout)

If dropout is random, not a problem

- Example of Baumannn, Seifert-Kessel, Jones (1992):

> Because of illness and transfer to another school, incomplete data were obtained for one subject each from the TA and DRTA group

Problematic drop-out or exclusion

If loss of units due to treatment or underlying conditions, problematic!

Rosensaal (2021) rebuking a study on the effectiveness of hydrochloriquine as treatment for Covid19 and reviewing allocation:

> Of these 26, six were excluded (and incorrectly labelled as lost to follow-up): three were transferred to the ICU, one died, and two terminated treatment or were discharged

Sick people excluded from the treatment group! then claim it is better.

Worst: "The index [treatment] group and control group were drawn from different centres."

Why seek balance?

Two main reasons

1. Power considerations: with equal variance in each group, balanced samples gives the best allocation
2. Simplicity of interpretation and calculations: the interpretation of the F test in a linear regression is unambiguous

Finding power in balance

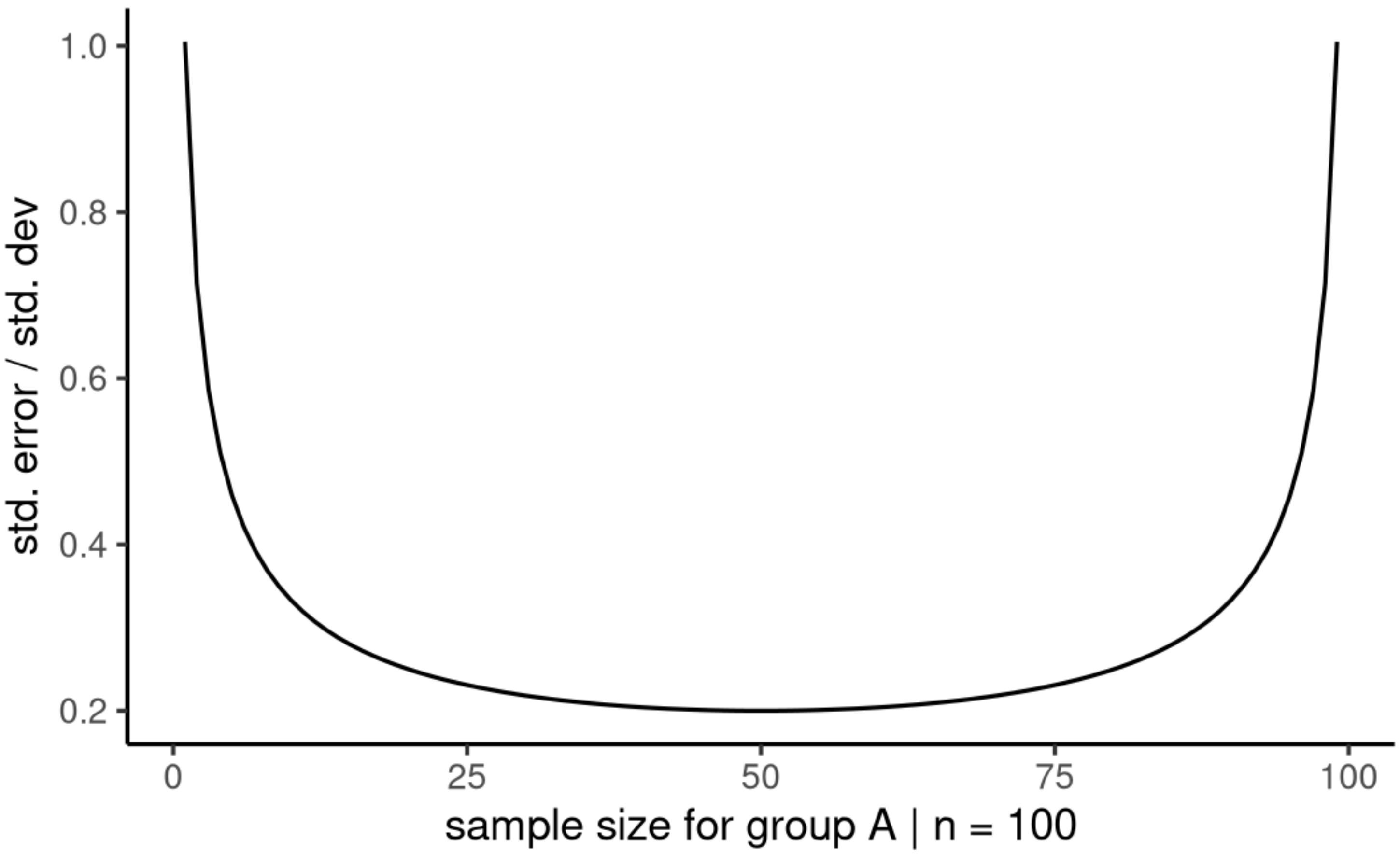
Consider a t-test for assessing the difference between treatments A and B with equal variability

$$t = \frac{\text{estimated difference}}{\text{estimated variability}} = \frac{(\widehat{\mu}_A - \widehat{\mu}_B) - 0}{\mathsf{se}(\widehat{\mu}_A - \widehat{\mu}_B)}.$$

The standard error of the average difference is

$$\sqrt{\frac{\text{std. dev}_A}{\text{nb of obs. in } A} + \frac{\text{std. dev}_B}{\text{nb of obs. in } B}} = \sqrt{\frac{\sigma}{n_A} + \frac{\sigma}{n_B}}$$

Optimal allocation of ressources



The allocation of $n = n_A + n_B$ units that minimizes the std error is $n_A = n_B = n/2$.

Example: tempting fate

We consider data from Multi Lab 2, a replication study that examined Risen and Gilovich (2008) who

> explored the belief that tempting fate increases bad outcomes. They tested whether people judge the likelihood of a negative outcome to be higher when they have imagined themselves [...] tempting fate [...] (by not reading before class) or not [tempting] fate (by coming to class prepared). Participants then estimated how likely it was that [they] would be called on by the professor (scale from 1, not at all likely, to 10, extremely likely).

The replication data gathered in 37 different labs focuses on a 2 by 2 factorial design with gender (male vs female) and condition (prepared vs unprepared) administered to undergraduates.

Load data | Check balance | Marginal means

```r
# This is a 2x2 factorial design
# The response is 'likelihod'
# the explanatories are 'condition' and 'gender'
library(tidyverse)
url1 <- "https://edsm.rbind.io/data/RG08rep.csv"
RS_unb <- read_csv(url1, col_types = c("iiff"))
# Data artificially balanced for the sake
# of illustration purposes
url2 <- "https://edsm.rbind.io/data/RG08rep_bal.csv"
RS_bal <- read_csv(url2, col_types = c("iiff"))
```

Load data | **Check balance** | Marginal means

```r
summary_stats <-
  RS_unb %>%
  group_by(condition) %>%
  summarize(nobs = n(),
            mean = mean(likelihood))
```

Summary statistics

condition	nobs	mean
unprepared	2192	4.606
prepared	2241	4.060

Load data | Check balance | **Marginal means**

```r
options(contrasts = c("contr.sum",
                      "contr.poly"))
model <- lm(likelihood ~ gender*condition
            data = RS_unb)
library(emmeans)
emm <- emmeans(model,
               specs = "condition")
```

Marginal means for condition

condition	emmean	SE
unprepared	4.504	0.0540
prepared	4.022	0.0535

Note unequal standard errors.

Explaining the discrepancies

Estimated marginal means are based on equiweighted groups:

$$\widehat{\mu} = \frac{1}{4}(\widehat{\mu}_{11} + \widehat{\mu}_{12} + \widehat{\mu}_{21} + \widehat{\mu}_{22})$$

where $\widehat{\mu}_{ij} = n_{ij}^{-1} \sum_{r=1}^{n_{ij}} y_{ijr}$.

The sample mean is the sum of observations divided by the sample size.

The two coincide when $n_{11} = \cdots = n_{22}$.

Why equal weight?

- The ANOVA and contrast analyses, in the case of unequal sample sizes, are generally based on marginal means (same weight for each subgroup)
- This choice is justified because research questions generally concern comparisons of means across experimental groups.

Revisiting the F statistic

Statistical tests contrast competing **nested** models:

- an alternative (full) model
- a null model, which imposes restrictions (a simplification of the alternative models)

The numerator of the F-statistic compares the sum of square of a model with (given) main effect, etc. to a model without.

What is explained by condition?

Consider the 2×2 factorial design with factors A: `gender` and B: `condition` (prepared vs unprepared) without interaction.

What is the share of variability (sum of squares) explained by the experimental condition?

Comparing differences in sum of squares (1)

Consider a balanced sample

```
anova(lm(likelihood ~ 1, data = RS_bal),
      lm(likelihood ~ condition, data = RS_bal))
# When gender is present
anova(lm(likelihood ~ gender, data = RS_bal),
      lm(likelihood ~ gender + condition, data = RS_bal))
```

The difference in sum of squares is 141.86 in both cases.

Comparing differences in sum of squares (2)

Consider an unbalanced sample

```
anova(lm(likelihood ~ 1, data = RS_unb),
      lm(likelihood ~ condition, data = RS_unb))
# When gender is present
anova(lm(likelihood ~ gender, data = RS_unb),
      lm(likelihood ~ gender + condition, data = RS_unb))
```

The differences of sum of squares are respectively 330.95 and 332.34.

Orthogonality

Balanced designs yield orthogonal factors: the improvement in the goodness of fit (characterized by change in sum of squares) is the same regardless of other factors.

So effect of B and $B \mid A$ (read B given A) is the same.

- test for $B \mid A$ compares $\mathsf{SS}(A, B) - \mathsf{SS}(A)$
- for balanced design, $\mathsf{SS}(A, B) = \mathsf{SS}(A) + \mathsf{SS}(B)$ (factorization).

We lose this property with unbalanced samples: there are distinct formulations of ANOVA.

Analysis of variance - Type I (sequential)

The default method in **R** with `anova` is the sequential decomposition: in the order of the variables A, B in the formula

- So F tests are for tests of effect of
  - A, based on $\mathsf{SS}(A)$
  - $B \mid A$, based on $\mathsf{SS}(A, B) - \mathsf{SS}(A)$
  - $AB \mid A, B$ based on $\mathsf{SS}(A, B, AB) - \mathsf{SS}(A, B)$

Ordering matters

Since the order in which we list the variable is **arbitrary**, these F tests are not of interest.

Analysis of variance - Type II

Impact of

- $A \mid B$ based on $\mathsf{SS}(A, B) - \mathsf{SS}(B)$
- $B \mid A$ based on $\mathsf{SS}(A, B) - \mathsf{SS}(A)$
- $AB \mid A, B$ based on $\mathsf{SS}(A, B, AB) - \mathsf{SS}(A, B)$
- tests invalid if there is an interaction.
- In **R**, use `car::Anova(model, type = 2)`

Analysis of variance - Type III

Most commonly used approach

- Improvement due to $A \mid B$, AB, $B \mid A$, AB and $AB \mid A$, B
- What is improved by adding a factor, interaction, etc. given the rest
- may require imposing equal mean for rows for $A \mid B$, AB, etc.
  - (**requires** sum-to-zero parametrization)
- valid in the presence of interaction
- but F-tests for main effects are not of interest
- In **R**, use `car::Anova(model, type = 3)`

ANOVA for unbalanced data

```r
model <-
  lm(likelihood ~ condition*gender,
     data = RS_unb)
# Three distinct decompositions
anova(model) #type 1
car::Anova(model, type = 2)
car::Anova(model, type = 3)
```

ANOVA (type I)

	Df	Sum Sq	F value
gender	1	164.94	29.1
condition	1	332.34	58.7
gender:condition	1	36.55	6.5
Residuals	4429	25086.33	

ANOVA (type II)

	Df	Sum Sq	F value
gender	1	166.33	29.4
condition	1	332.34	58.7
gender:condition	1	36.55	6.5
Residuals	4429	25086.33	

ANOVA (type III)

	Df	Sum Sq	F value
gender	1	167.71	29.6
condition	1	227.88	40.2
gender:condition	1	36.55	6.5
Residuals	4429	25086.33	

ANOVA for balanced data

```r
model2 <-
  lm(likelihood ~ condition*gender,
     data = RS_bal)
anova(model2) #type 1
car::Anova(model2, type = 2)
car::Anova(model2, type = 3)
# Same answer - orthogonal!
```

ANOVA (type I)

	Df	Sum Sq	F value
condition	1	141.86	24.1
gender	1	121.69	20.6
condition:gender	1	37.88	6.4
Residuals	2500	14733.84	

ANOVA (type II)

	Df	Sum Sq	F value
condition	1	141.86	24.1
gender	1	121.69	20.6
condition:gender	1	37.88	6.4
Residuals	2500	14733.84	

ANOVA (type III)

	Df	Sum Sq	F value
condition	1	141.86	24.1
gender	1	121.69	20.6
condition:gender	1	37.88	6.4
Residuals	2500	14733.84	

Recap

- If each observation has the same variability, a balanced sample maximizes power.
- Balanced designs have interesting properties:
  - estimated marginal means coincide with (sub)samples averages
  - the tests of effects are unambiguous
  - for unbalanced samples, we work with marginal means and type 3 ANOVA
  - if empty cells (no one assigned to a combination of treatment), cannot estimate corresponding coefficients (typically higher order interactions)

Practice

From the OSC psychology replication

> People can be influenced by the prior consideration of a numerical anchor when forming numerical judgments. [...] The anchor provides an initial starting point from which estimates are adjusted, and a large body of research demonstrates that adjustment is usually insufficient, leading estimates to be biased towards the initial anchor.

Replication of Study 4a of Janiszewski & Uy (2008, Psychological Science) by J. Chandler

Polynomial regression

IJLR: It's Just a Linear Regression...

All ANOVA models we covered so fall (t-tests, factorial designs, latin squares) are all special instances of the linear regression model.

The latter says that

$$\underbrace{\mathsf{E}(Y_i)}_{\text{average response}} = \underbrace{\beta_0 + \beta_1 \mathbf{X}_{1i} + \cdots + \beta_p \mathbf{X}_{pi}}_{\text{linear (i.e., additive) combination of explanatories}}$$

What about factors?

The software eats **numbers**, not labels.

What happens under the hood with the sum-to-zero constraint?

Assuming that level a of factor A does not appear in the coefficient table, including A requires adding $(a-1)$ vectors $\mathbf{X}_j$ where

$$\mathrm{X}_{ij} = \begin{cases} 1 & A = j, \\ -1 & A = a, \\ 0 & \text{otherwise.} \end{cases}$$

Check `model.matrix()` on a linear model object in **R**.

Beyond ANOVA

Consider linear model with a single **continuous** explanatory, where X is an experimental factor.

We assume that $Y_i \sim \mathsf{No}\{\text{smooth function}(\mathrm{X}_i), \sigma^2\}$.

Approximate the smooth function of X by a pth order polynomial,

$$\mathsf{E}(Y_i) = \beta_0 + \beta_1 \mathrm{X}_i + \cdots + \beta_p \mathrm{X}_i^p$$

Example: Bean soaking

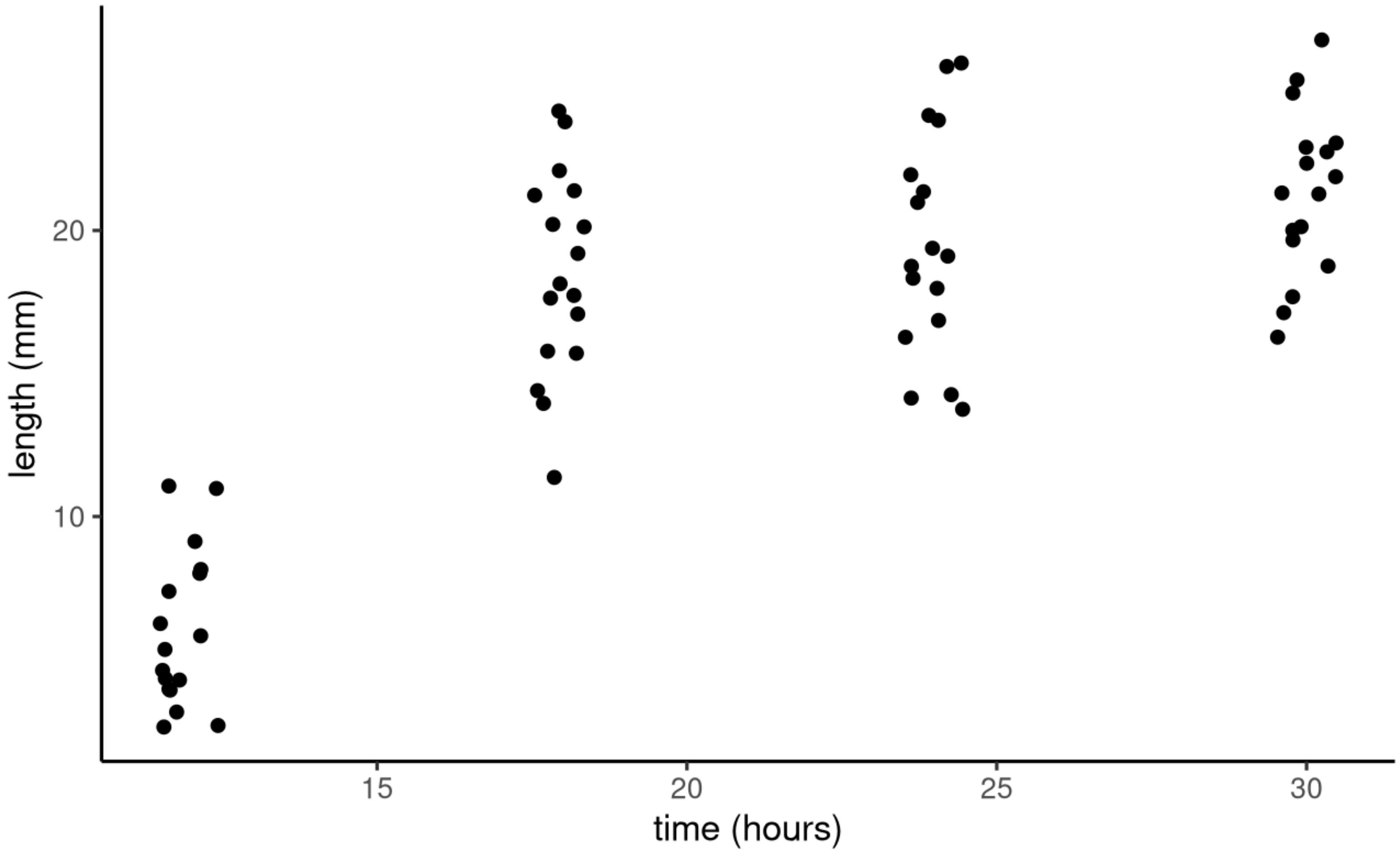
Example 8.8 of Dean, Voss and Draguljić

> What is the optimal soaking time of beans prior to planting?

Experimental factor: time (in hours), either 12, 18, 24 and 30 hours (equally spaced).

```
url <- "https://edsm.rbind.io/data/bean.txt"
beans <- read.table(url, header = TRUE)
g1 <- ggplot(data = beans,
       aes(x = time, y = length)) +
    geom_point(position = position_jitter(width = 0.5)) +
  theme_classic() +
  labs(y = 'length (mm)',
       x = 'time (hours)')
```

Beans data



Soaking times are jittered to avoid overplotting.

Trend model or ANOVA?

Fitting the cubic model is equivalent to a one-way ANOVA with time (four levels) with $r = 17$ replications.

In each case, there are four parameters. For time $\texttt{time} \in \{12, 18, 24, 30\}$ hours associated to level j of the categorical variable:

$$\mathsf{E}(\texttt{length}) = \mu + \alpha_j = \beta_0 + \beta_1 \texttt{time} + \beta_2 \texttt{time}^2 + \beta_3 \texttt{time}^3.$$

The difference is that we cannot interpolate with the one-way ANOVA for times between 12 and 30.

Testing for higher-order terms

Test nested models using F tests: null $\subset$ alternative

In the model

$$\mathsf{E}(\texttt{length}) = \beta_0 + \beta_1\texttt{time} + \beta_2\texttt{time}^2 + \beta_3\texttt{time}^3$$

- $\mathscr{H}_0 : \beta_3 = 0$, the coefficient associated to the cubic term $\texttt{time}^3$.
- $\mathscr{H}_0 : \beta_2 = \beta_3 = 0$, compare cubic vs linear model.

Fitting polynomials in **R**

```
model3 <- lm(length ~ poly(time, degree = 3),
             data = beans) #cubic model
model2 <- lm(length ~ poly(time, degree = 2),
             data = beans) #quadratic
model1 <- lm(length ~ poly(time, degree = 1),
             data = beans) #linear
model_anov <- lm(length ~ factor(time),
                 data = beans) #one-way ANOVA
```

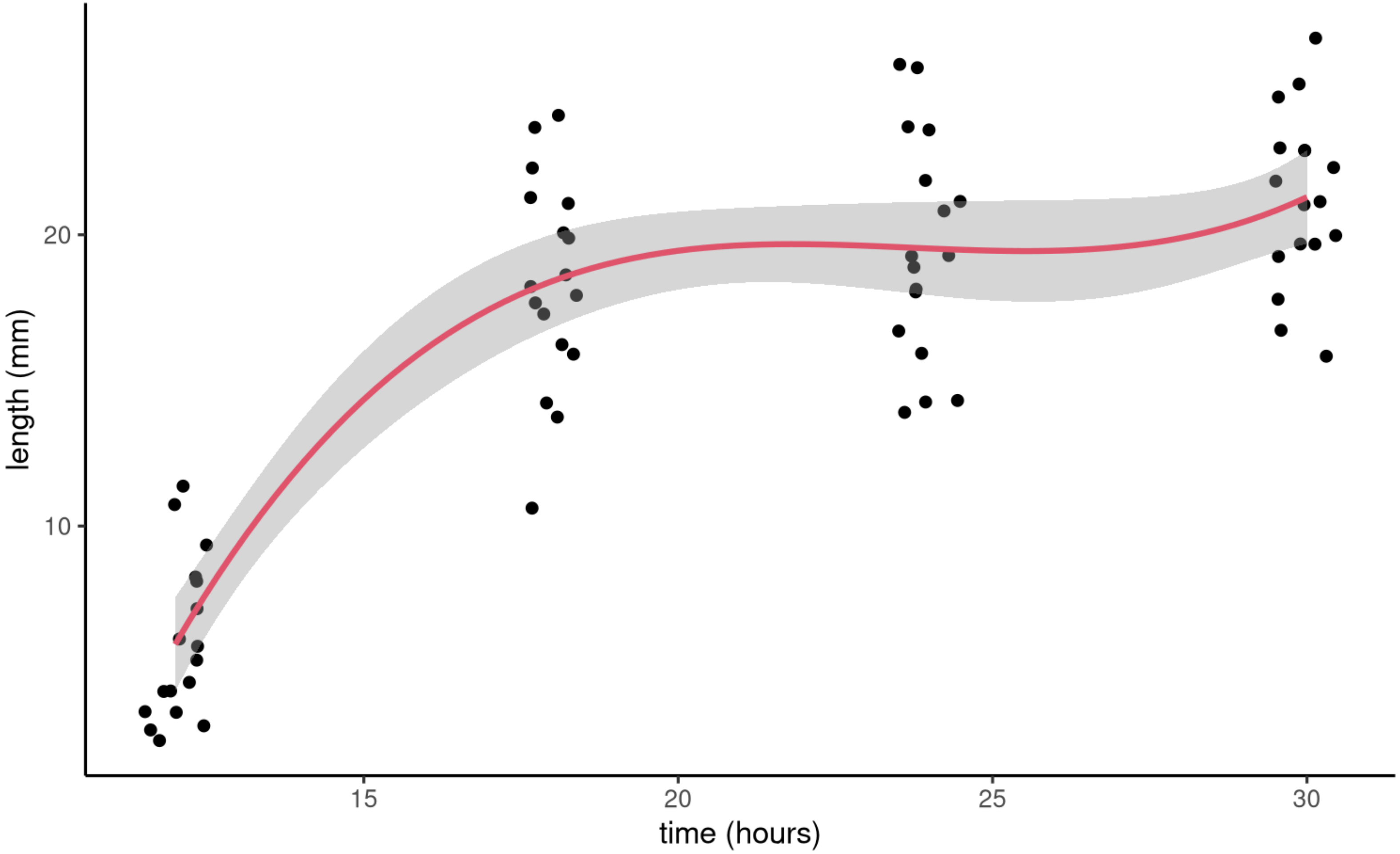
The function `poly` uses orthogonal polynomial (more stable numerically).

Model comparisons (F tests)

```
# Model 3 is equivalent to ANOVA
anova(model3, model_anov)
# drop cubic term?
anova(model2, model3) #H0: beta3=0
# drop quadratic + cubic?
anova(model1, model3) #H0: beta2 = beta3=0
```

We cannot simplify the cubic model: p-value less than 0.001407.

Fitted model



Pairwise comparisons

Compute pairwise differences with Tukey's method

```r
library(emmeans)
pairwise_diff <-
 contrast(
  emmeans(model_anov,
          specs = "time"),
  method = "pairwise",
  adjust = "tukey",
  level = 0.99,
  infer = c(TRUE, FALSE))
```

Pairwise differences with 99% CI (Tukey's method)

contrast	difference	lower CI	upper CI
12 - 18	-12.47	-16.15	-8.79
12 - 24	-13.59	-17.27	-9.91
12 - 30	-15.35	-19.04	-11.67
18 - 24	-1.12	-4.80	2.57
18 - 30	-2.88	-6.57	0.80
24 - 30	-1.76	-5.45	1.92

Every soaking time is significantly better than 12 hours