Lecture 2: Normal Linear Models — Basics

MATH3823 Generalised Linear Models

Richard P Mann

MATH3823 Generalised Linear Models

Reading

Course notes: Chapter 2, Sections 2.1–2.2

`www.richardpmann.com/MATH3823`

Motivating Example: Birth Weight Data

Dataset: Birth weights of 24 babies (12 girls, 12 boys)

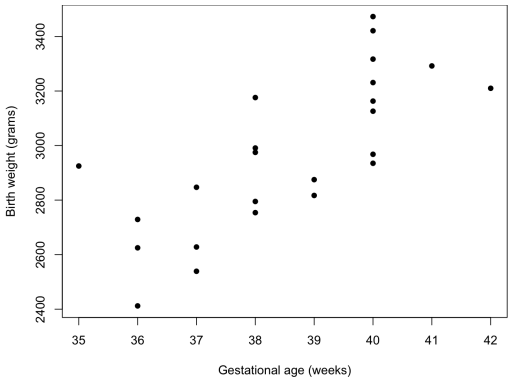
Variables:

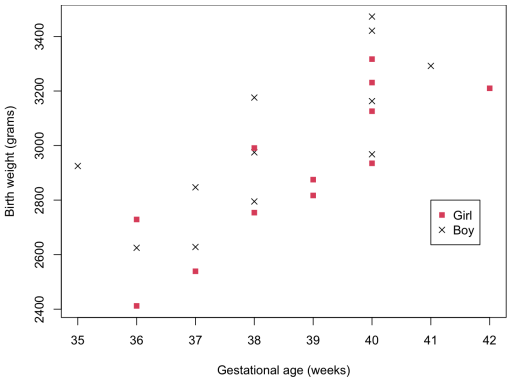
- **Response:** Birth weight (grams)
- **Explanatory:**
  - Gestational age (continuous, in weeks)
  - Sex (categorical: 0 = girl, 1 = boy)

Questions:

- How does birth weight depend on gestational age?
- Do boys and girls differ in birth weight?
- Is the effect of age different for boys and girls?

Exploring the Data





Clear positive relationship between gestational age and birth weight. Boys (orange) appear slightly heavier than girls (blue) at similar ages.

Four Nested Models

Model 0: Constant only

$$\text{Weight} = \alpha$$

Model 1: Age effect only

$$\text{Weight} = \alpha + \beta \cdot \text{Age}$$

Model 2: Parallel lines (age + sex, no interaction)

$$\text{Weight} = \alpha + \beta \cdot \text{Age} + \gamma \cdot \text{Sex}$$

Model 3: Interaction (separate slopes)

$$\text{Weight} = \alpha + \beta \cdot \text{Age} + \gamma \cdot \text{Sex} + \delta \cdot \text{Age} \times \text{Sex}$$

Terminology

Response variable: The outcome we want to predict (birth weight)

Explanatory variables: Predictors/covariates (age, sex)

Parameters: Unknown quantities to estimate ($\alpha, \beta, \gamma, \delta$)

Main effects: Individual variable contributions (β, γ)

Interaction: When effect of one variable depends on another (δ)

Interaction interpretation:

- If $\delta \neq 0$: The effect of age on weight differs between boys and girls
- If $\delta = 0$: Age has the same effect regardless of sex

Residual Sum of Squares

For any model M_k with r_k residual degrees of freedom:

$$R_k = \sum_{i=1}^{n} (y_i - \hat{\mu}_{ki})^2$$

where $\hat{\mu}_{ki}$ is the fitted value for observation i under model M_k.

Model	Parameters	Residual df
M_0: Constant	1	$n-1$
M_1: Age	2	$n-2$
M_2: Age + Sex	3	$n-3$
M_3: Age $\times$ Sex	4	$n-4$

Comparing Nested Models: The F-Test

To compare nested models $M_0 \subset M_1$:

$$F_{01} = \frac{(R_0 - R_1)/(r_0 - r_1)}{R_1/r_1}$$

Under H_0: M_0 is adequate (simpler model is sufficient):

$$F_{01} \sim F_{r_0 - r_1, r_1}$$

Decision rule:

- Large $F \Rightarrow$ Small p-value $\Rightarrow$ Reject $H_0 \Rightarrow$ Prefer M_1
- Small $F \Rightarrow$ Large p-value $\Rightarrow$ Don't reject $H_0 \Rightarrow$ Use simpler M_0

Birth Weight: Model 1 Results

Fitted model: Weight $= \hat{\alpha} + \hat{\beta} \times$ Age

	Estimate	Std. Error	t value	p-value
Intercept	-1484.98	833.85	-1.78	0.089
Age	115.53	22.10	5.23	< 0.001

Test: Model 0 vs Model 1

$$F_{01} = 27.33 \sim F_{1,22}, \quad p = 3.04 \times 10^{-5}$$

$\Rightarrow$ Strong evidence that age affects birth weight.

Birth Weight: Model 2 Results

Fitted model: Weight $= \hat{\alpha} + \hat{\beta} \times$ Age $+ \hat{\gamma} \times$ Sex

	Estimate	Std. Error	t value	p-value
Intercept	-1773.32	820.91	-2.16	0.042
Age	120.89	21.42	5.64	< 0.001
Sex	163.04	72.81	2.24	0.036

Test: Model 1 vs Model 2

$$F_{12} = 5.01 \sim F_{1,21}, \quad p = 0.036$$

$\Rightarrow$ Evidence that boys are heavier than girls (by $\approx$ 163g on average).

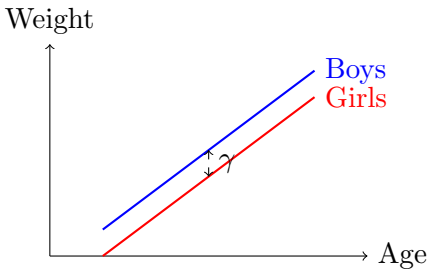
Interpreting the Parallel Lines Model

For girls (Sex = 0):

$$\mathbb{E}[\text{Weight}] = -1773 + 121 \times \text{Age}$$

For boys (Sex = 1):

$$\mathbb{E}[\text{Weight}] = (-1773 + 163) + 121 \times \text{Age} = -1610 + 121 \times \text{Age}$$



Same slope (121 g/week), different intercepts.

Dummy Variables for Categorical Predictors

How to include Sex in the model?

Create a **dummy variable** (indicator variable):

$$\text{Sex}_i = \begin{cases} 0 & \text{if observation } i \text{ is female} \\ 1 & \text{if observation } i \text{ is male} \end{cases}$$

Interpretation of coefficient γ:

- γ = difference in mean response between males and females
- Females are the **reference category** (baseline)
- Males are compared *to* females

Note: Choice of reference category is arbitrary but affects interpretation.

What About Interaction?

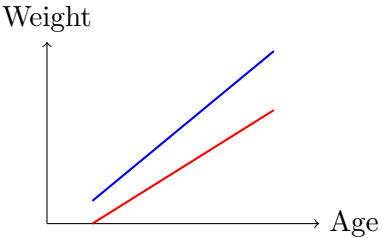
Model 3: Weight $= \alpha + \beta \times$ Age $+ \gamma \times$ Sex $+ \delta \times$ Age $\times$ Sex

Test: Model 2 vs Model 3

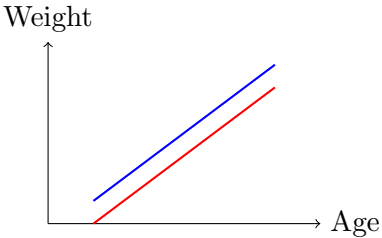
$$F_{23} \sim F_{1,20}$$

If p-value is large $\Rightarrow$ No evidence for interaction.

Implication: The effect of gestational age on birth weight is the same for boys and girls.



Interaction: different slopes



No interaction: parallel

Model Fitting in R

```r
# Load data
birthwt <- read.csv("birthweight.csv")

# Fit models
model0 <- lm(weight ~ 1, data = birthwt)
model1 <- lm(weight ~ age, data = birthwt)
model2 <- lm(weight ~ age + sex, data = birthwt)
model3 <- lm(weight ~ age * sex, data = birthwt)

# Compare models
anova(model0, model1)  # Test age effect
anova(model1, model2)  # Test sex effect
anova(model2, model3)  # Test interaction

# Summary of chosen model
summary(model2)
```

R Output: Key Components

From `summary(model2)`:

- **Coefficients table:** Estimates, standard errors, t-values, p-values
- **Residual standard error:** $\hat{\sigma}$ (estimate of error SD)
- **Multiple R-squared:** Proportion of variance explained
- **F-statistic:** Overall model significance (vs. null model)

From `anova(model1, model2)`:

- **RSS:** Residual sum of squares for each model
- **Df:** Difference in degrees of freedom
- **F:** F-statistic for nested comparison
- **Pr(>F):** p-value

Summary

Key points:

- Normal linear models assume $Y \mid \mathbf{x} \sim \mathcal{N}(\mu(\mathbf{x}), \sigma^2)$
- Nested models can be compared using F-tests
- Categorical variables enter via dummy variables
- The reference category is set to 0
- Interactions allow slopes to differ across groups
- R's `lm()` function fits normal linear models

Next lecture: Matrix formulation and types of linear models.