Course 2 · Week 1 — Linear models end-to-end

Cheatsheet — biostats_courses

R. Heller

Correlation vs regression

- Correlation: symmetric; both variables random (**Model II**).
- Regression: asymmetric; predictor fixed, outcome random (**Model I**).
- Use Model II (e.g. `smatr::sma`) when both variables have error.

Simple linear regression

$y_i = \beta_0 + \beta_1 x_i + \varepsilon_i, \quad \varepsilon_i \sim N(0, \sigma^2)$

```
fit <- lm(y ~ x, data = df)
broom::tidy(fit, conf.int = TRUE)
broom::glance(fit)
```

Multiple regression

```
fit <- lm(y ~ x1 + x2 + x1:x2, data = df)
```

Concept	Fix
Confounding	adjust by including the confounder
Interaction	include `x1:x2` (or `x1 * x2`)
Non-linearity	`splines::ns(x, 4)` or `I(x^2)`
Scale sensitivity	centre / standardise predictors

Diagnostics

```
plot(fit)                    # base: 4 diagnostic plots
performance::check_model(fit)   # tidy overview
car::vif(fit)                # multicollinearity
```

Plot	What it flags
Residuals vs fitted	non-linearity, non-constant variance
Normal Q-Q	non-normal residuals
Scale-location	heteroscedasticity
Leverage / Cook's	influential outliers

VIF > 5 ⟶ investigate collinearity; > 10 ⟶ fix it.

Robust / weighted

```
MASS::rlm(y ~ x, data = df)          # robust to outliers
lm(y ~ x, weights = 1 / var_i)        # weighted LS
```

```r
# Heteroscedasticity-consistent SEs
library(sandwich); library(lmtest)
coeftest(fit, vcov = vcovHC(fit, type = "HC3"))
```

Decision rule for Week 1

- Residual plot ugly? Transform, add terms, or switch link.
- VIF explodes? Drop one of the collinear predictors or combine them.
- Outlier with high Cook's distance? Investigate before deleting.
- Variance depends on X? HC3 SEs or weighted LS.

Common pitfalls

- Interpreting β_1 as "effect of X" under confounding.
- Reporting R^2 on in-sample data as if it were generalisable.
- Including an interaction but only reporting main effects.
- Centring categorical predictors (don't — use reference coding).

Further reading

- Harrell, *Regression Modeling Strategies*, ch. 4–5.
- Fox, *Applied Regression Analysis*.