Clinical Regression in R

Choose the model, encode the estimand, diagnose the fit, report the effect

Rverse Analytics

The outcome distribution chooses the model; the scientific question chooses the contrast. Write the estimand before fitting anything: adjusted mean difference, odds ratio, rate ratio or hazard ratio?

Model map

Outcome	Starting model	Core R call	Effect scale
Continuous	Gaussian linear	`lm(y ~ x + z, data = d)`	adjusted mean difference
Binary	Logistic	`glm(y ~ x + z, family = binomial, data = d)`	OR after `exp(coef)`
Common binary risk	Modified Poisson + robust SE	`glm(y ~ x + z, family = poisson("log"), data = d)`	risk ratio
Count / person-time	Poisson	`glm(events ~ x + offset(log(time)), family = poisson, data = d)`	incidence-rate ratio
Overdispersed count	Negative binomial	`MASS::glm.nb(events ~ x + offset(log(time)), data = d)`	incidence-rate ratio
Ordered category	Proportional odds	`ordinal::clm(y ~ x + z, data = d)`	common OR
Repeated / clustered	Mixed or GEE	`lme4::glmer(...)` / `geepack::geeglm(...)`	conditional / marginal
Time-to-event	Cox PH	`survival::coxph(Surv(t, event) ~ x + z, data = d)`	hazard ratio

A reproducible fit-to-table pattern

```
library(dplyr)
library(broom)
library(gtsummary)

d <- trial |>
  mutate(
    response = factor(response, levels = c(0, 1), labels = c("No", "Yes")),
    stage = relevel(stage, ref = "T1")
  )

fit <- glm(response ~ scale(age) + stage + grade,
           family = binomial, data = d, na.action = na.exclude)

# Machine-readable estimates
tidy(fit, exponentiate = TRUE, conf.int = TRUE)

# Publication table
tbl_regression(fit, exponentiate = TRUE,
  label = list(`scale(age)` ~ "Age, per 1 SD")) |>
  add_global_p() |>
  add_n() |>
  bold_labels()
```

Encode predictors deliberately

```
# Reference category
d$group <- relevel(factor(d$group), ref = "Control")

# Non-linearity: restricted cubic spline
fit_spline <- glm(y ~ splines::ns(age, df = 4) + sex,
                  family = binomial, data = d)

# Prespecified interaction: effect of treatment differs by sex
fit_int <- glm(y ~ treatment * sex + age,
               family = binomial, data = d)

# Likelihood-ratio test for the interaction block
anova(update(fit_int, . ~ . - treatment:sex), fit_int, test = "LRT")
```

Interpretation: With an interaction, the main treatment coefficient is the treatment effect at the reference level of `sex`; it is not the overall treatment effect.

Diagnostics by model

Question	Useful check	R recipe	Red flag / response
Linearity	residuals vs fitted	`plot(fit, which = 1)`	curve → transform/spline
Constant variance	scale-location	`plot(fit, which = 3)`	fan shape → robust SE/model variance
Influential cases	Cook's distance	`plot(fit, which = 4)`	investigate data and sensitivity
Collinearity	VIF	`performance::check_collinearity(fit)`	unstable CI → revise predictors
Logistic calibration	observed vs predicted	`performance::check_model(fit)`	miscalibration → respecify/validate
Discrimination	ROC/AUC	`pROC::roc(d$y, fitted(fit))`	report CI; AUC is not calibration
Count dispersion	Pearson χ^2 / df	`sum(residuals(fit, "pearson")^2) / df.residual(fit)`	>1 → NB or robust approach

Robust uncertainty and model comparison

```r
# Heteroskedasticity-consistent covariance
V <- sandwich::vcovHC(fit, type = "HC3")
lmtest::coeftest(fit, vcov. = V)

# Nested models: likelihood-ratio test
anova(fit_small, fit_large, test = "LRT")

# Predict on the response scale with uncertainty
pred <- predict(fit, newdata = new_patients, type = "link", se.fit = TRUE)
new_patients |>
  mutate(prob = plogis(pred$fit),
         lo = plogis(pred$fit - 1.96 * pred$se.fit),
         hi = plogis(pred$fit + 1.96 * pred$se.fit))
```

Report, do not merely print

- State population, outcome coding, predictors, functional forms and missing-data strategy.
- Report effect estimate + 95% CI + exact *p*; include units and reference categories.
- Separate prespecified adjustment from data-driven selection; avoid univariable *p*-value screening.
- Give events and candidate parameter counts; quantify optimism with internal validation when predicting.
- For prediction models, report calibration and discrimination; a significant coefficient is not validation.
