

8

Regression Modelling: Logistic, and Beyond

Linear Regression as a Foundation

Linear regression models a continuous outcome as a linear function of one or more predictors plus random error, and it is the conceptual foundation for nearly every regression-based method that follows in this book, including logistic and Poisson regression (generalised linear models with a different link function and error distribution, Section 8.3) and the hazard regression models of Chapter 9. Fitted by ordinary least squares, the model estimates how the expected outcome changes per unit change in a predictor, holding other predictors constant — for example, estimating grams of birthweight gained per additional antenatal care visit, adjusting for maternal age and parity.

Four assumptions must be checked, not merely assumed: linearity of the relationship, independence of errors, homoscedasticity (constant error variance), and approximate normality of residuals (important mainly for small samples, since the Central Limit Theorem provides some robustness in larger samples). Violations are diagnosed through residual plots, and remedies include transforming the outcome (e.g., log-transforming a skewed cost or length-of-stay variable) or moving to a generalised linear model with an appropriate distribution.

Model Building, Confounding, and Interaction

A central purpose of multivariable regression in health research is confounder adjustment: isolating the association between an exposure and an outcome from the influence of other variables associated with both. A classic Nigerian example is the observed association between distance to the nearest health facility and stunting prevalence, which is confounded by household wealth (poorer households are both further from facilities on average and at higher nutritional risk); a regression model that adjusts for wealth quintile isolates the distance effect more credibly than a bivariate comparison. Interaction terms extend this further, allowing an association to differ by a third variable — for example, testing whether the effect of insecticide-treated net (ITN) ownership on malaria incidence differs between rural and urban LGAs — and interpreting an interaction correctly requires reporting the association separately within each level of the moderating variable, not merely the significance of the interaction term.

Logistic Regression for Binary Outcomes

When the outcome is binary — a positive/negative diagnostic result, survival/death, treatment success/failure — logistic regression models the log-odds of the outcome as a linear function of predictors, and exponentiating a coefficient yields an odds ratio, interpreted as the multiplicative change in the odds of the outcome per unit change in the predictor, holding other variables constant. Logistic regression is the workhorse model behind clinical risk scores and behind much of the malaria and NTD risk-factor literature in Nigeria — for example, modelling the odds of a confirmed positive malaria test as a function of age, season, ITN use, and state, or modelling the odds of severe adolescent depressive symptomatology as a function of household food security, school attendance, and gender.

R example

```
# R: logistic regression for malaria positivity, with odds ratios and 95% CIs
model <- glm(confirmed_positive ~ age_group + season + itn_use + state,
             family = binomial(link = "logit"), data = patient_level_data)

exp(cbind(OR = coef(model), confint(model))) # odds ratios with confidence intervals
```

Poisson and Negative Binomial Regression for Count Outcomes

Count outcomes — monthly confirmed malaria cases per facility, number of leishmaniasis cases per LGA per year — are naturally modelled with Poisson regression, which assumes the mean equals the variance. Health surveillance data very commonly violate this assumption (overdispersion: variance exceeds the mean, often because of unmeasured heterogeneity across facilities or clustering of cases in outbreaks), and the negative binomial regression model, which adds a dispersion parameter, is the standard remedy; a simple diagnostic (comparing the residual deviance to its degrees of freedom) indicates whether overdispersion is present and a switch is warranted. An offset term (the log of the population at risk, or the number of tests performed) is essential whenever comparing counts across facilities or LGAs of different sizes, converting the model from one of raw counts to one of rates.

Model Evaluation and Selection

A fitted regression model should be evaluated on multiple fronts: goodness of fit (deviance, pseudo-R-squared for logistic/Poisson models), discrimination for classification models (the area under the ROC curve, or AUC, quantifying how well the model separates positive from negative cases), calibration (whether predicted probabilities match observed frequencies — a model can have excellent discrimination but poor calibration, which matters

greatly if predicted probabilities will directly inform a clinical or resource-allocation decision), and out-of-sample validation (performance on data not used to fit the model, ideally from a different time period or a different set of facilities, to guard against overfitting — a theme developed formally in Chapter 11).

Worked Example

Interpreting a Full Logistic Regression Output

Consider a logistic regression predicting confirmed malaria positivity with the following fitted odds ratios: age 5-14 vs. under-5 (OR = 1.35, 95% CI 1.05-1.73), rainy vs. dry season (OR = 2.80, 95% CI 2.20-3.56), ITN ownership vs. none (OR = 0.62, 95% CI 0.49-0.79), and urban vs. rural residence (OR = 0.71, 95% CI 0.55-0.92). Interpreting each: children aged 5-14 have 35% higher odds of a positive test than under-5s, holding season, ITN use, and residence constant; testing in the rainy season is associated with 180% higher odds of positivity than the dry season; owning an ITN is associated with 38% lower odds of positivity; and urban residents have 29% lower odds of positivity than rural residents. Because all four confidence intervals exclude 1, all four associations are statistically significant at the conventional 0.05 level. Note carefully that these are odds ratios, not risk ratios or probabilities directly — a common source of misinterpretation when odds ratios are casually described as ‘X times more likely.’

Predictor	Odds ratio	95% CI	Interpretation
Age 5-14 vs. <5	1.35	1.05-1.73	35% higher odds
Rainy vs. dry season	2.80	2.20-3.56	180% higher odds
ITN owner vs. none	0.62	0.49-0.79	38% lower odds
Urban vs. rural	0.71	0.55-0.92	29% lower odds

Worked Example

Diagnosing and Correcting Overdispersion

A Poisson regression of monthly leishmaniasis case counts on year and state produces a residual deviance of 812 on 340 degrees of freedom — a ratio of roughly 2.4, well above the value close to 1 expected if the Poisson assumption held, indicating substantial overdispersion. Refitting the same model specification as a negative binomial regression (which estimates an additional dispersion parameter) yields the same direction and similar magnitude of coefficients but noticeably wider standard errors and confidence intervals — a more honest reflection of the true uncertainty, since the Poisson model's artificially narrow intervals would have overstated the precision of the estimated trend.

R example

```
# R: diagnosing overdispersion and switching to negative binomial regression
library(MASS)

poisson_model <- glm(cases ~ year + state + offset(log(population)),
  family = poisson(link = "log"), data = nli_annual)
cat("Deviance / df:", deviance(poisson_model) / df.residual(poisson_model), "\n")

nb_model <- glm.nb(cases ~ year + state + offset(log(population)), data = nli_annual)
summary(nb_model)$coefficients # compare standard errors to the Poisson model above
```

PYTHON example

```
# Python: the same overdispersion check and negative binomial fit with statsmodels
import statsmodels.api as sm
import statsmodels.formula.api as smf

poisson_model = smf.glm("cases ~ year + state", data=nli_annual,
  family=sm.families.Poisson(),
  offset=np.log(nli_annual["population"])).fit()
print("Deviance / df:", poisson_model.deviance / poisson_model.df_resid)

nb_model = smf.glm("cases ~ year + state", data=nli_annual,
  family=sm.families.NegativeBinomial(),
  offset=np.log(nli_annual["population"])).fit()
print(nb_model.summary())
```

Worked Example

ROC Curves and Calibration Together

A logistic model for severe malaria risk at triage achieves an AUC of 0.81, indicating good discrimination (a randomly chosen severe case is ranked higher-risk than a randomly chosen non-severe case 81% of the time). However, a calibration plot (observed proportion of severe outcomes within each decile of predicted risk, plotted against the mean predicted risk in that decile) shows the model systematically under-predicts risk in the highest-risk decile — predicting an average of 45% risk where the observed rate is actually 62%. This combination (good discrimination, poor calibration at the extremes) is common and consequential: a clinician using the raw predicted probability to counsel a caregiver in the highest-risk group would be given a falsely reassuring number, even though the model correctly ranks that patient as high-risk relative to others.

Figure 8.1 — ROC Curve for the Severe Malaria Triage Model (Sec. 8.8)

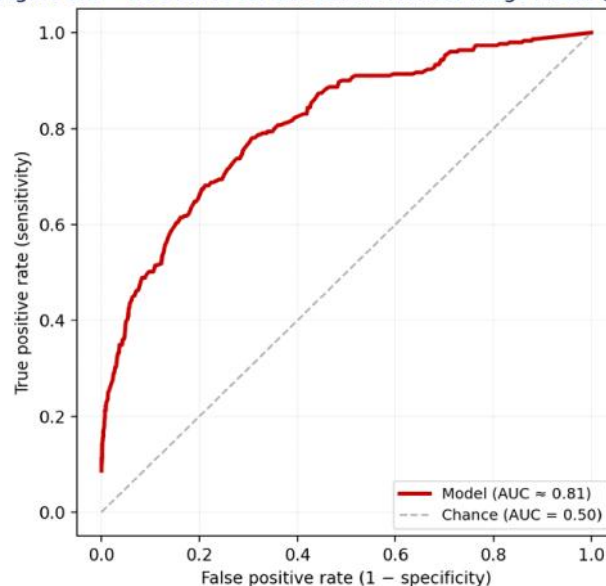


Figure 8.1 — ROC curve for the severe malaria triage model (Section 8.8).

Reporting a Regression Model: A Checklist

- State the outcome, all predictors, and the model type (linear/logistic/Poisson) explicitly, including how any categorical variable's reference category was chosen.
- Report the sample size and the number of observations excluded due to missing data, with a brief statement of how missingness was handled (Chapter 4).
- Report effect estimates (coefficients, odds ratios, or rate ratios) with 95% confidence intervals, not p-values alone.
- Report at least one goodness-of-fit or discrimination measure appropriate to the model type.
- State explicitly which associations are being interpreted causally (if any) and what assumptions (e.g., no unmeasured confounding) that interpretation requires.

Extended Case Study**A Model That Looked Right But Was Reversed**

A cautionary and instructive real-world pattern: an analyst fits a logistic regression predicting severe malaria outcome as a function of, among other variables, whether the patient had received a blood transfusion during their hospital stay, and finds a striking and highly statistically significant result — transfusion is associated with roughly 6 times higher odds of the severe outcome. A hurried interpretation might report this as evidence that transfusion is somehow harmful or a marker of poor-quality care. The correct interpretation, however, requires attending carefully to the direction of causality and the timing of variable measurement relative to the outcome, a point this chapter's discussion of confounding (Section 8.2) touches on but which deserves fuller emphasis here: patients receive a blood transfusion specifically **BECAUSE** they are already severely anaemic, which is itself a component of, or strong risk factor for, severe malaria — so

transfusion is not a cause of severe outcome at all, but rather a downstream consequence and marker of it, a form of what is sometimes called ‘reverse causation’ or, more precisely here, treating a variable measured during or after the outcome’s development as if it were a baseline predictor of that outcome.

The remedy is not a different statistical technique but a more careful specification of the model: any variable that is only measured, or that can only take a meaningful value, after the outcome of interest has already begun to develop should be excluded from a model intended to predict severe outcome at the point of triage, since it could not actually have been available to inform a real-time clinical decision at that point, and its strong statistical association reflects the outcome’s own progression rather than a genuine, actionable risk factor. This distinction — between predictors genuinely available at the moment a prediction needs to be made, and variables that are merely correlated with the outcome because they occur alongside or after it — is one of the most common and consequential errors in applied predictive modelling, in health data science and beyond, and checking the timing of every candidate predictor relative to the outcome should be a standard, explicit step (not an afterthought) in any model-building workflow, alongside the more familiar checks for statistical confounding.

Standardised Coefficients and Comparing Predictor Importance

A common question when presenting a multivariable regression result is which predictor ‘matters most’ — a question the raw, unstandardised coefficients cannot directly answer, because predictors measured on very different scales (age in years versus a symptom score from 0-27) produce coefficients that are not directly comparable in magnitude. Standardising each continuous predictor (subtracting its mean and dividing by its standard

deviation before fitting the model) puts all coefficients on a common scale, expressing each one as the expected change in the outcome per one-standard-deviation change in that predictor, allowing a much fairer visual and numerical comparison of relative importance across predictors measured in entirely different original units — though this comparison should still be interpreted alongside domain knowledge, since a predictor’s statistical standardised effect size is not the same as its practical or clinical importance.

Mixed-Effects Models

Regression for Grouped Data

The regression models presented in this chapter assume, by default, that all observations are independent — an assumption Section 3.8 and Section 11.3 have both already flagged as frequently violated in health data with a natural grouping structure (patients within facilities, repeated measures within patients). Mixed-effects (multilevel/hierarchical) regression models extend the logistic and Poisson regression framework of this chapter specifically to handle this structure, adding a ‘random effect’ term that allows the model to explicitly represent facility-to-facility (or patient-to-patient) variation in the baseline outcome rate, rather than forcing a single overall intercept onto data that plausibly varies systematically by group. Fitting a mixed-effects model instead of a standard regression on facility-clustered data (using, for example, R’s ‘lme4’ package or Python’s ‘statsmodels’ mixed linear model implementation) typically produces appropriately wider, more honest standard errors for facility-level or patient-level predictors, directly analogous to the corrected, design-based standard errors introduced for survey data in Section 7.4, and is the standard, recommended approach whenever a regression model’s data has this kind of natural multilevel grouping structure.

Model Comparison Criteria

AIC, BIC, and Their Appropriate Use

When comparing several candidate regression models that are not simple nested subsets of each other, information criteria such as the Akaike Information Criterion (AIC) and Bayesian Information Criterion (BIC) provide a principled way to balance model fit against complexity, penalising the addition of predictors that do not sufficiently improve the model's fit to justify their inclusion; a lower AIC or BIC indicates a preferred model, with BIC generally imposing a heavier penalty for additional parameters and therefore favouring more parsimonious models than AIC when the two disagree. These criteria are best used to compare a small, substantively motivated set of candidate models specified in advance, rather than as the sole basis for an automated, exhaustive search across every possible combination of predictors, which reintroduces exactly the multiple-comparisons and overfitting risks discussed in Chapters 7 and 11.

Quantile Regression

Beyond the Conditional Mean

Ordinary linear regression (Section 8.1) models how the mean of an outcome changes with a predictor, but the mean is not always the quantity of greatest interest — a nutrition programme concerned specifically with the most severely undernourished children may care more about how a predictor affects the 10th percentile of a weight-for-age z-score distribution than about its effect on the average child. Quantile regression directly models a specified quantile (or several quantiles simultaneously) of the outcome's conditional distribution, and comparing coefficients across quantiles can reveal that a predictor's effect is not constant across the distribution — for example, that household food insecurity has a much larger negative association with weight-for-age at the lower tail of the distribution (the most vulnerable

children) than at the median, a pattern an ordinary mean-focused regression would entirely miss.

A Complete Regression Diagnostic Function

The regression checklist in Section 8.9 lists what a complete report should contain; encoding that checklist directly into a function ensures it is actually applied every time a model is fitted, rather than remaining a checklist an analyst might forget to consult under deadline pressure. The function below fits a logistic regression and returns, in one call, the coefficients as odds ratios with confidence intervals, the AUC, and a basic overdispersion-style check — directly operationalising Sections 8.5 through 8.9 as reusable code rather than a manual, easily-skipped routine.

PYTHON example

```
# Python: a regression fitting function that returns a complete diagnostic bundle
import statsmodels.api as sm
from sklearn.metrics import roc_auc_score

def fit_logistic_with_diagnostics(X, y):
    X_const = sm.add_constant(X)
    model = sm.Logit(y, X_const).fit(dis=0)
    odds_ratios = model.params.apply(lambda b: round(2.718281828 ** b, 3))
    ci = model.conf_int().apply(lambda col: (2.718281828 ** col).round(3))
    predicted = model.predict(X_const)
    auc = roc_auc_score(y, predicted)
    return {"odds_ratios": odds_ratios.to_dict(), "confidence_intervals": ci.to_dict(),
            "auc": round(auc, 3), "n_obs": int(model.nobs)}

diagnostics = fit_logistic_with_diagnostics(X, y)
print(diagnostics) # one call now returns everything Section 8.9's checklist requires
```

Chapter Summary

- Linear regression's assumptions (linearity, independence, homoscedasticity, normal residuals) should be checked via residual diagnostics, not assumed.
- Multivariable regression adjusts for confounding; interaction terms allow associations to vary by a moderating variable and must be interpreted stratum-by-stratum.

- Logistic regression models binary outcomes via log-odds; exponentiated coefficients are odds ratios, which must not be casually described as risk ratios.
- Poisson regression models counts; overdispersion (common in surveillance data) calls for negative binomial regression, and an offset converts counts to rates.
- Model evaluation should cover fit, discrimination (AUC), calibration, and out-of-sample validation — a model can discriminate well while still being poorly calibrated, especially at the extremes of risk.
- A full regression report interprets every coefficient's direction, magnitude, and confidence interval in plain language, not just its p-value.

Mini-Project

Mini-Project 8: Building and Fully Reporting a Regression Model

Fit an appropriate regression model (linear, logistic, or Poisson/negative binomial) to a dataset of your choice and produce a complete, publication-style results table and interpretation.

Step 1. Choose an outcome and at least three predictors, including at least one you expect to confound the relationship between another predictor and the outcome.

Step 2. Fit the model without and then with the potential confounder, and compare how the coefficient of interest changes.

Step 3. If your outcome is binary or a count, check the appropriate diagnostic (calibration for logistic; deviance/df for Poisson) and correct if necessary.

Step 4. Produce a results table (following the format of Section 8.6) with effect estimates, confidence intervals, and plain-language interpretations.

Step 5. Write a short paragraph distinguishing which of your findings are both statistically and practically significant.

Deliverable: *A results table, diagnostic check output, and a half-page interpretation.*

End-of-Chapter Exercises

1. A linear regression of birthweight on maternal ANC visits shows a funnel-shaped residual plot. Name the assumption being violated and one remedy.

2. Interpret an odds ratio of 1.8 for ITN use (coded 1 = owns net) predicting malaria-negative test result — including the direction of coding.
3. A Poisson regression of monthly case counts shows residual deviance far exceeding its degrees of freedom. What does this suggest, and what model would you fit instead?
4. Using the worked example in Section 8.6, explain why saying ‘children aged 5-14 are 35% more likely to test positive’ is an imprecise restatement of the actual odds ratio finding, and give a more accurate phrasing.
5. In Section 8.7, explain why the negative binomial model’s standard errors are wider than the Poisson model’s, even though the point estimates are similar.
6. A model has an AUC of 0.85 but is poorly calibrated in the highest-risk decile (Section 8.8). Explain, for a specific clinical use case of your choosing, why this combination could still be dangerous in practice.
7. Fit (or describe how you would fit) an interaction between ITN use and residence (urban/rural) in a logistic model of malaria positivity, and explain what a statistically significant interaction term would mean in this context.
8. Using the checklist in Section 8.9, identify one item you most often see omitted from regression results you have read, and explain why its omission is a problem.
9. Explain why stating the reference category for a categorical predictor is essential for correctly interpreting a regression coefficient.
10. A published regression table reports only p-values with no confidence intervals or sample size. Using Section 8.9, list what additional information you would request from the authors.

11. In the Section 8.10 case study, explain why blood transfusion's very strong statistical association with severe outcome does not mean transfusion is a genuine risk factor for it.
12. Describe the general rule you would apply, when building a predictive model, to decide whether a candidate predictor should be excluded for the reason illustrated in Section 8.10.
13. Give one other example (real or plausible) of a health variable that could show a similarly misleading 'reverse causation' pattern in a predictive model.
14. Why is this problem described as being about predictor TIMING rather than about confounding in the sense of Section 8.2?
15. If a variable's timing relative to the outcome is genuinely ambiguous or unknown, what would you do before including it in a triage risk model?
16. Using Section 8.11, explain why comparing the raw (unstandardised) coefficients for age (in years) and a 0-27 symptom score directly would be misleading.
17. Describe, in your own words, how standardising a predictor changes the interpretation of its regression coefficient.
18. Why does the chapter caution that a standardised effect size is 'not the same as' practical or clinical importance?
19. Propose a scenario where a predictor has a large standardised coefficient but low practical importance, or vice versa.
20. Would you recommend reporting standardised coefficients, raw coefficients, or both, in a report intended for a mixed clinical and statistical audience? Justify your choice.

21. Using Section 8.12, explain why a standard logistic regression on facility-clustered patient data could understate the true uncertainty around a facility-level predictor's effect.
22. What does the 'random effect' term in a mixed-effects model represent, in your own words?
23. Explain the connection between mixed-effects models (Section 8.12) and the design-based survey standard errors introduced in Section 7.4.
24. Propose a dataset from earlier in this book (e.g., the malaria diagnostics or adolescent cohort data) where a mixed-effects model would be more appropriate than a standard regression, and explain why.
25. Why might ignoring a genuine facility-level clustering structure lead to an overconfident, rather than merely imprecise, regression result?
26. Summarise, in your own words, why this chapter treats confounding, reverse causation, and clustering as three distinct threats to a regression model's validity.
27. Explain the common thread linking the reverse-causation case study (Section 8.10) and the mixed-effects models discussion (Section 8.12).
28. Why does this chapter insist on checking model assumptions and reporting a full results table, rather than just quoting a headline coefficient?
29. Propose a regression model from your own field where you would need to check specifically for the reverse-causation pattern illustrated in Section 8.10.
30. If you had to give one piece of advice to someone about to build their first logistic regression model, what would it be?

31. Using Section 8.14, explain why a nutrition programme focused on the most severely undernourished children might prefer quantile regression over ordinary linear regression.
32. Describe a scenario where a predictor's effect could differ substantially between the 10th percentile and the median of an outcome's distribution.
33. Why would an ordinary mean-focused regression miss the pattern quantile regression is designed to detect?
34. Propose one health outcome, other than weight-for-age, where the tail of the distribution might be more important than the mean.
35. What practical cost (in terms of interpretability or complexity) comes with fitting quantile regression across several quantiles rather than a single mean regression?