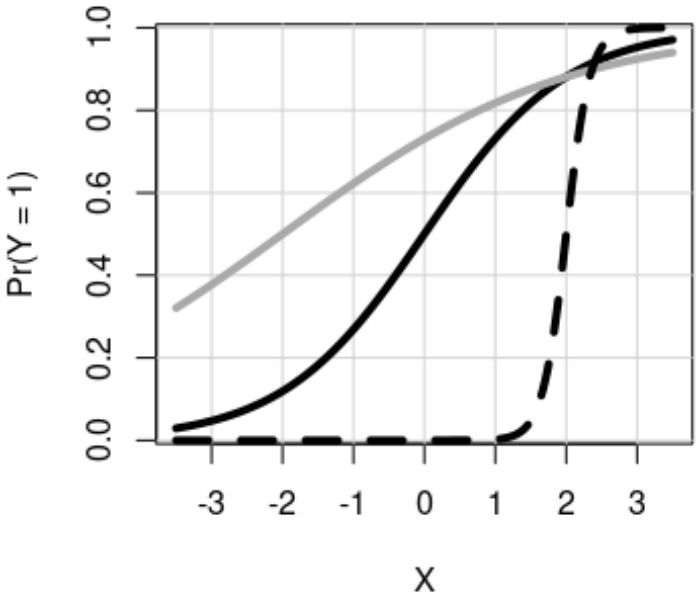
Chapter 10: Models for binary dependent variables

1. The figure below shows three inverse-logit curves on the same set of axes. The equations for these curves are

    - $P(Y = 1) = \text{logit}^{-1}(X)$
    - $P(Y = 1) = \text{logit}^{-1}\left(1 + \frac{1}{2}X\right)$
    - $P(Y = 1) = \text{logit}^{-1}(-12 + 6X)$



Answer the following:

- Match each equation to its graphical representation in the plot.
- For each curve state the intercept and the slope on the log-odds scale.
- Which curve implies the largest change in probability for a one-unit increase from $X = 0$ to $X = 1$? Briefly explain why.

2. Load the Titanic passenger data used in the chapter, using `read.csv("titanic.csv")`. In this exercise we will extend the analysis presented in the chapter to investigate whether the association between age and survival differs by passenger class.

- *Before* fitting anything, sketch (mentally or on paper) the full set of fixed-effect coefficients you expect to appear when you add the interaction `age_centred:passengerClass` to the model. How many *additional* coefficients will `summary()` print?
- Next, fit the model with the interaction term, and state whether there is statistical evidence that the age–survival slope differs between at least two passenger classes. – Report the estimated age slope (log-odds per extra year) in first class and in third class.

    - Translate the third-class slope into an odds ratio for a 10-year age difference.
    - Briefly describe, in plain English, how the effect of age on survival changes with ticket class.

3. The `wells.csv` contains data from a study in which researchers measured arsenic levels in wells in an area of Bangladesh and indicated whether it was safe to drink[1] or contaminated with arsenic (containing more than 0.5 in units of hundreds of micrograms per liter). Households using unsafe wells were encouraged to switch to the nearest safe well. Place it in your working directory, then read it into R using `read.csv("wells.csv")`.

   The data contains these variables

   - `switch`: 1 = switched, 0 = did not switch
   - `dist`: distance to the closest safe well (metres)
   - `arsenic`: arsenic concentration in the current well (×100 µg/L)
   - `education`: years of schooling of the household head
   - `association`: participation in local community organisation (as a 0/1 indicator variable) by any member of the household.
   - Fit a logistic regression with distance as the sole predictor:
     ▪ Is dist a statistically significant predictor of switching?
     ▪ Report the odds of switching when dist = 100.
     ▪ How does that odds change for each additional metre of distance?
     ▪ Convert the odds in part 2 to a probability.
   - Refit the model including arsenic level. Does higher arsenic make switching more likely?

       - Which of these two households is more likely to have switched wells?
         • Household A has an unsafe well with arsenic level of 1.8, and the

[1] This dataset is taken from the excellent book *Data Analysis Using Regression and Multilevel/Hierarchical Models* by Andrew Gelman and Jennifer Hill. Analysis of these data was previously published in Gelman, A., Trevisani, M., Lu, H. and Van Geen, A. (2004), 'Direct Data Manipulation for Local Decision Analysis as Applied to the Problem of Arsenic in Drinking Water from Tube Wells in Bangladesh.' *Risk Analysis,* 24: 1597-1612.

closest safe well is at a distance of 350 metres.
• Household B has an unsafe well with arsenic level of 0.6, and the closest safe well is at 200 metres of distance.

- Extend the model with education and association. Does participation in community organisations (association) significantly influence switching?
    - Is education a significant predictor?
    - How does the odds of switching wells changes with each additional years of education of the household's head?
    - Suppose that, based on our model, a certain household C has a probability of 0.5 of switching wells. You discover that the head of the household actually has two more years of education than recorded. Without rerunning the model, update the probability estimate. *Hint: work on the log-odds scale and then transform back*.

4. One long-standing question in decision-making research is how well people exploit their internal sense of confidence – the feeling that a just-made choice was probably correct – to guide subsequent choices. One approach to study this in the lab is the *dual-decision task*[2], in which every trial contains two successive two-alternative forced-choice (2AFC) decisions:

Stage	Stimulus rule	Implication
Decision 1	Either option is correct with 50% probability (chance level).	Choose whichever option looks more likely.
Decision 2	Which option is correct *depends on whether Decision 1 was right* (if the first choice was correct, option A is correct; otherwise, option B is).	Knowing you were right should boost accuracy.

An optimal decision-maker who perfectly tracks their own correctness should therefore be more accurate on Decision 2 than on Decision 1. A subset of the data from this study is included in the file `dual_decision.csv`, which contains three columns:

- `accuracy` – 1 = correct, 0 = incorrect
- `decision` – `"1st"` / `"2nd"`
- `id` – participant code (26 participants)

- Inspect the data. Load the `.csv` file and verify the structure and counts for each decision stage.

[2] Lisi, M., Mongillo, G., Milne, G., Dekker, T. & Gorea, A. (2021) 'Discrete confidence levels revealed by sequential decisions.' *Nature Human Behaviour, 5*, 273–280. https://doi.org/10.1038/s41562-020-00953-1

- Fit a multilevel logistic regression model, predicting `accuracy` as a function of `decision`. Include both by-participant random intercepts and random slopes.
- Does accuracy differ between the two stages? Identify the fixed-effect coefficient that tests this, and report its estimate, the standard error, z-value and p-value. State whether the null hypothesis (no difference) can be rejected.
- Exponentiate the `decision2nd` coefficient to obtain the odds ratio (OR) – the ratio of the odds of a correct response on Decision 2 versus Decision 1. Interpret this OR in plain language (e.g., 'participants are *x* times more likely to be correct on Decision 2') and compute a 95 per cent confidence interval around it (e.g. using the method illustrated in Chapter 9 for linear multilevel models).
- Using the fixed-effect estimates, compute the model-predicted probability of a correct response for an 'average' participant at each stage. Report both probabilities and the absolute gain in percentage points.
- *(Optional visual check)* For each participant, plot their observed proportion correct on Decision 2 (y-axis) against their proportion correct on Decision 1 (x-axis). Add the identity line $y = x$. If most points lie above the diagonal, accuracy is indeed higher on the second decision.