Statistical Modeling, Winter 2025

Homework 5

Please submit your answers to these questions via Canvas prior to class next Tuesday. I prefer a pdf, or a .Rmd / .Qmd / .R file. Please name the file with your last name and hw number (e.g., elahi_hw1).

Complete the following:

- read chapter 10.2 and 11.1 in *Statistical Rethinking* (McElreath 2020).
- watch Modeling Events

Questions

From McElreath

1. If an event has probability 0.35, what are the log-odds of this event?

2. If an event has log-odds of 3.2, what is the probability of this event?

3. Create a vector of evenly spaced probabilities between 0 and 1 (but not including 0 and 1) in R. Then create another vector of the corresponding log-odds. Plot probability (y-axis) against log-odds (x-axis). Interpret the figure. What is the relevance of this figure to generalized linear modeling (e.g., logistic regression)?

4. Suppose that a coefficient in a logistic regression has a value of 1.7. What does this imply about the proportional change in odds of the outcome?

5. Finish the midterm. You can submit it as a separate file, or at minimum submit the following in this HW doc:

- one prior predictive distribution justifying your choice of priors
- diagnose your chains
- compare WAIC for the models
- interpret your results

Project

1. What is your response variable? Which of the probability distributions in the exponential family (see Fig 10.6 in SR2) is most appropriate to model your response? Why?

Answers

We start with a probability, p, for an event. The corresponding *odds* for that event are $p/(1-p)$. The *log-odds* are thus given by:

$$\operatorname{logit}(p) = \log \frac{p}{1-p}$$

Using this information, we can answer question 1.

1

```r
# Define logit function
logit <- function(p) log(p / (1 - p))
# 1. log-odds of p = 0.35
logit(0.35)
```

```
[1] -0.6190392
```

What if we want to go in the reverse direction, that is, from a log-odds to a probability?

If we let $y = \operatorname{logit}(p)$ and do some algebra, we can derive an expression for the $inverse - logit$, also known as the *logistic*:

$$\begin{aligned}
y &= \log \frac{p}{1-p} \\
\mathrm{e}^y &= \frac{p}{1-p} \\
\mathrm{e}^y(1-p) &= p \\
\mathrm{e}^y - \mathrm{e}^y p &= p \\
\mathrm{e}^y &= p + \mathrm{e}^y p \\
\mathrm{e}^y &= p(1+\mathrm{e}^y) \\
p &= \frac{\mathrm{e}^y}{1+\mathrm{e}^y}
\end{aligned}$$

Now we can write an inverse-logit function and answer problem 2.

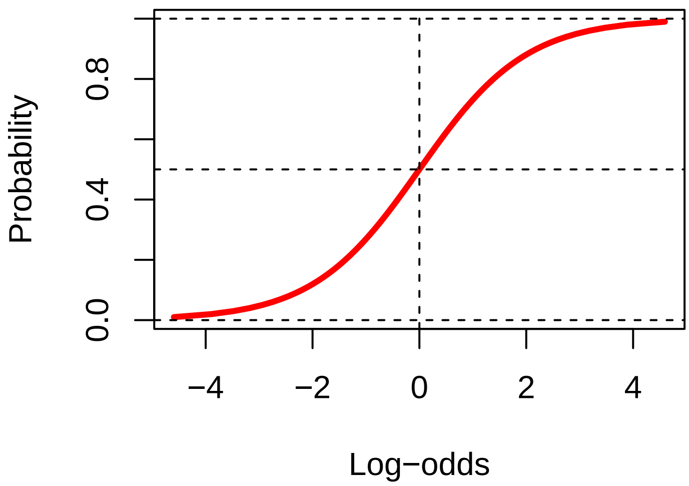
2

```r
# Define inverse logit function
inv_logit <- function(x) (exp(x)) / (1 + exp(x))
# 2. prob of log-odds = 3.2
inv_logit(3.2)
```

```
[1] 0.9608343
```

3

```
p_grid <- seq(0.01, 0.99, by = 0.01)
log_odds <- logit(p_grid)
plot(y = p_grid, x = log_odds, type = "l", lwd = 3,
     xlab = "Log-odds", ylab = "Probability", col = "red")
abline(h = c(0,0.5,1), lty = 2)
abline(v = 0, lty = 2)
```



If we were to model probability with a normal distribution, we would run into problems - the golem would predict values outside of the *support* of the *random variable*. To get around this, we can convert probability to log-odds and then use a normal distribution (because as indicated in the plot above, log-odds are not bounded).

However, it is conventional to use a *logit link* for a binomial regression and model p directly (rather than transforming p first, and then modeling it using a normal). So the logistic function (i.e., the logit link) does the reverse - it maps the continuous linear model value to a probability parameter that is bounded between zero and one. The value of this non-intuitive approach is that we can model our response variable (counts) with an appropriate probability distribution (binomial). Enter *generalized* linear models.

4

Proportional odds is calculated by exponentiating the coefficient in a binomial GLM. The coefficient represents the *logodds* change for a unit change in the predictor, x.

```
exp(1.7)
```

```
[1] 5.473947
```

This means that each unit change in x multiplies the odds of the event by 5.5.

To demystify this relationship a little, if the linear model L is the log-odds of the event, then the odds of the event are $\exp(L)$. Now we want to compare the odds before and after increasing x by one unit. We want to know how much the odds increase, as a result of the unit increase in x. We can use algebra to solve this problem:

$$e^{(\alpha+\beta x)}Z = e^{(\alpha+\beta(x+1))}$$

The left side is the odds of the event, before increasing x. The Z represents the proportional change in odds that we're going to solve for. It's unknown value will make the left side equal to the right side. The right side is the odds of the event, after increasing x by 1 unit. Now we solve for Z:

$$\begin{aligned}
e^{(\alpha+\beta x)}Z &= e^{(\alpha+\beta(x+1))} \\
Z &= \frac{e^{(\alpha+\beta(x+1))}}{e^{(\alpha+\beta x)}} \\
&= e^{(\alpha+\beta(x+1)-(\alpha+\beta x))} \\
&= e^{(\alpha+\beta x+\beta-\alpha-\beta x))} \\
&= e^{\beta}
\end{aligned}$$

The answer is $Z = e^{\beta}$.

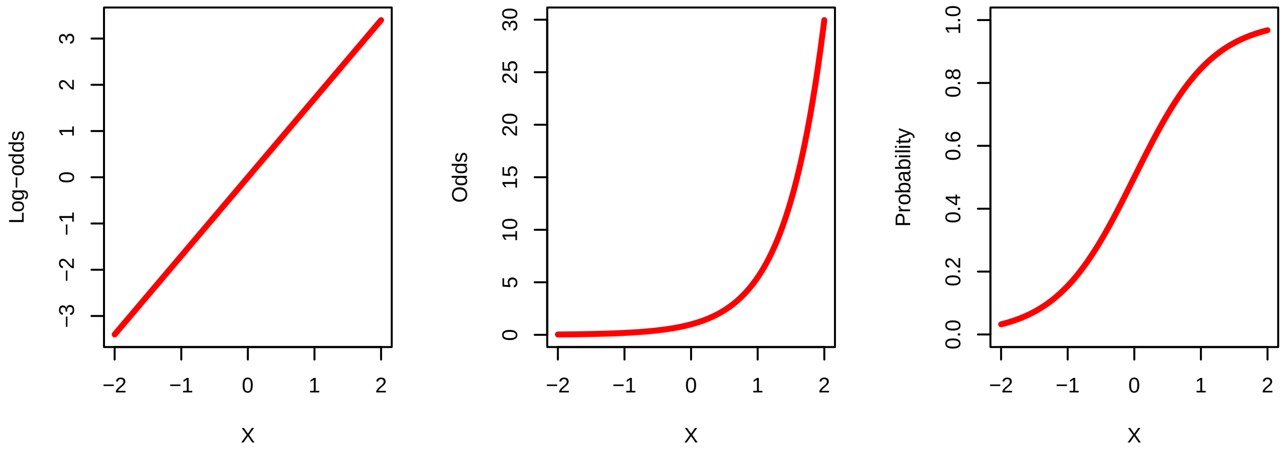
Let's visualize this coefficient:

```r
x <- seq(-2, 2, by = 0.01)
b <- 1.7
a <- 0
y <- a + b*x
set.seed(22)
y_prob <- inv_logit(y)
y_odds <- exp(y)

par(mfrow = c(1,3))
plot(x = x, y = y, type = "l", lwd = 3,
     xlab = "X", ylab = "Log-odds", col = "red")

plot(x = x, y = y_odds, type = "l", lwd = 3,
     xlab = "X", ylab = "Odds", col = "red")

plot(x = x, y = y_prob, type = "l", lwd = 3,
     xlab = "X", ylab = "Probability", col = "red",
     ylim = c(0, 1))
```



Check your math, so that the odds are actually being multiplied by the expected amount:

```
# Check math
df <- data.frame(x, y, y_odds, y_prob)

# Go from x=0, to x=1
df[x == 0, ]; df[x == 1, ]
```

```
    x y y_odds y_prob
201 0 0      1    0.5
```

```
    x   y   y_odds    y_prob
301 1 1.7 5.473947 0.8455347
```

```
df[x == 1, ]$y_odds / df[x == 0, ]$y_odds
```

```
[1] 5.473947
```

```
# Go from x=-1, to x=0
df[x == -0.5, ]; df[x == 0.5, ]
```

```
       x     y    y_odds    y_prob
151 -0.5 -0.85 0.4274149 0.2994329
```

```
      x    y   y_odds    y_prob
251 0.5 0.85 2.339647 0.7005671
```

```
df[x == 0.5, ]$y_odds / df[x == -0.5, ]$y_odds
```

```
[1] 5.473947
```

Play around with changing β and α - how does this change the curves?

References

McElreath, Richard. 2020. *Statistical Rethinking: A Bayesian Course with Examples in R and Stan*. CRC Press.