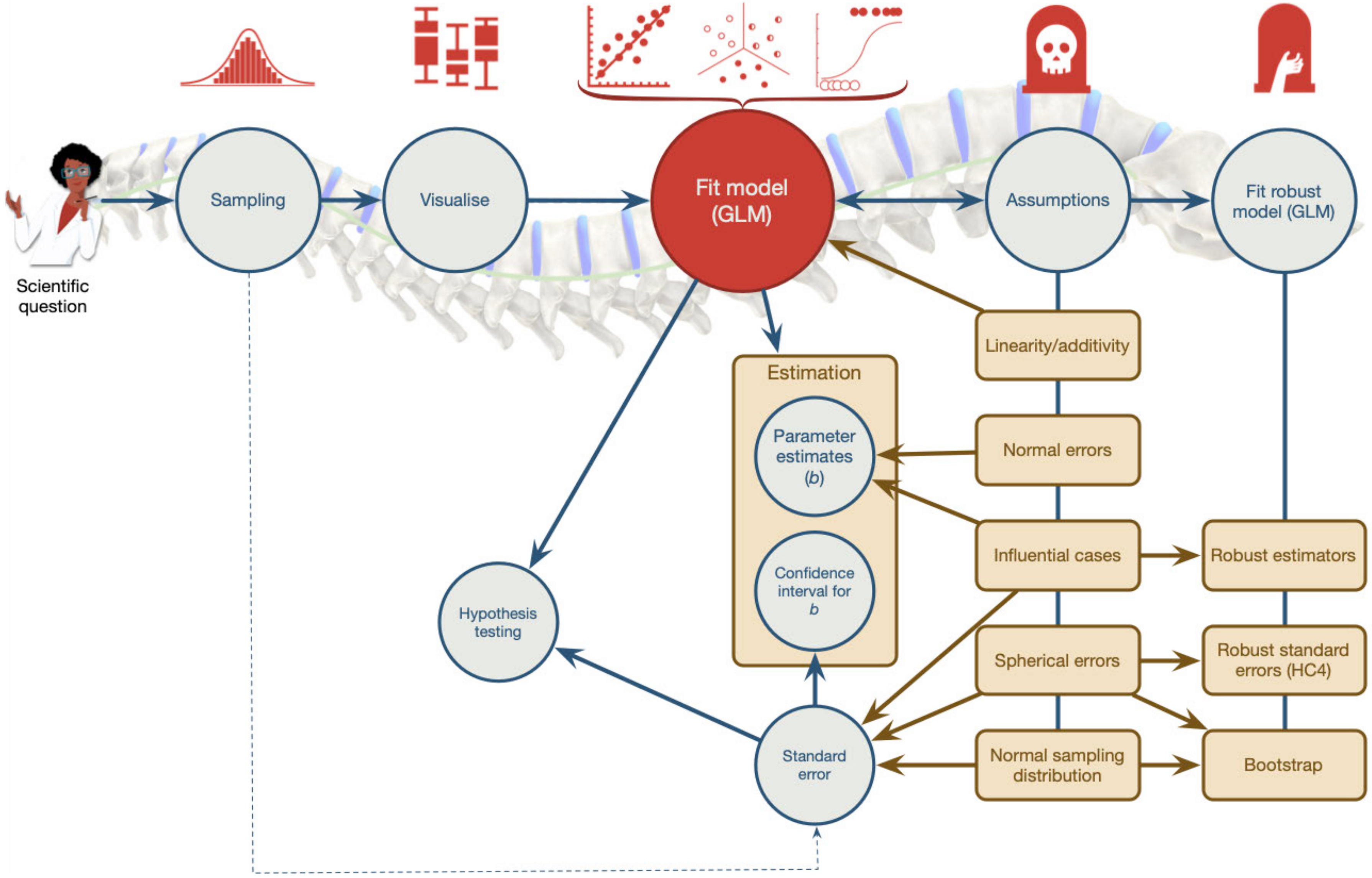
Categorical outcomes: logistic regression

Professor Andy Field

@profandyfield
www.youtube.com/user/ProfAndyField/
www.discoveringstatistics.com
www.milton-the-cat.rocks
www.discovr.rocks



Scientific question
Sampling
Visualise
Fit model (GLM)
Assumptions
Fit robust model (GLM)
Estimation
Parameter estimates (b)
Confidence interval for b
Hypothesis testing
Standard error
Linearity/additivity
Normal errors
Influential cases
Spherical errors
Normal sampling distribution
Robust estimators
Robust standard errors (HC4)
Bootstrap

A festive example

Santa Claus wanted to test the effects of different types of treats on whether presents got delivered:

- Predictors
  - treat: Christmas pudding, Mulled wine
- Outcome
  - delivered: Did the presents get delivered?

The data in R

Table 1: Santa's data

	id	treat	delivered
1	Fankle the Determined	Mulled wine	Not delivered
2	Amber the Cuddle	Mulled wine	Not delivered
3	Snorklum the Content	Pudding	Delivered
4	Henri the Biddible	Pudding	Delivered
5	Ainsley the Eigen vector	Mulled wine	Delivered
6	Funnelcup the Iron maiden fan	Mulled wine	Delivered
7	Ramsey the Surprised	Pudding	Delivered
8	Bindlestiff the Surprised	Mulled wine	Not delivered
9	Lardy the Strong	Pudding	Delivered
10	Theo the Small	Pudding	Delivered

Previous 1 2 3 4 5 … 40 Next

If it were a standard linear model

$$\text{delivered}_i = \hat{b}_0 + \hat{b}_1 \text{treat}_i + e_i$$

Assumption of linearity

- Violated with categorical outcomes
- We can't fit this model

Group	Treat
Christmas pudding	0
Mulled wine	1

But it's not a standard linear model

We predict the probability of the outcome occurring

$$P(Y) = \frac{1}{1 + e^{-(\hat{b}_0 + \hat{b}_1 X_i + e_i)}}$$

$$P(\text{delivery}) = \frac{1}{1 + e^{-(\hat{b}_0 + \hat{b}_1 \text{treat}_i + e_i)}}$$

Note the equation contains the linear model

Or ...

$$\ln\left(\frac{P(Y)}{1-P(Y)}\right) = \hat{b}_0 + \hat{b}_1 X_i + e_i$$

$$\ln\left(\frac{P(\text{delivery})}{1-P(\text{delivery})}\right) = \hat{b}_0 + \hat{b}_1 \text{treat}_i + e_i$$

Or ...

$$\ln\left(\frac{P(Y)}{1 - P(Y)}\right) = \hat{b}_0 + \hat{b}_1 X_i + e_i$$

$$\ln\left(\frac{P(\text{delivery})}{1 - P(\text{delivery})}\right) = \hat{b}_0 + \hat{b}_1 \text{treat}_i + e_i$$

Outcome

- We predict the log odds of the outcome occurring

Or ...

$$\ln\left(\frac{P(Y)}{1-P(Y)}\right) = \hat{b}_0 + \hat{b}_1 X_i + e_i$$

$$\ln\left(\frac{P(\text{delivery})}{1-P(\text{delivery})}\right) = \hat{b}_0 + \hat{b}_1 \text{treat}_i + e_i$$

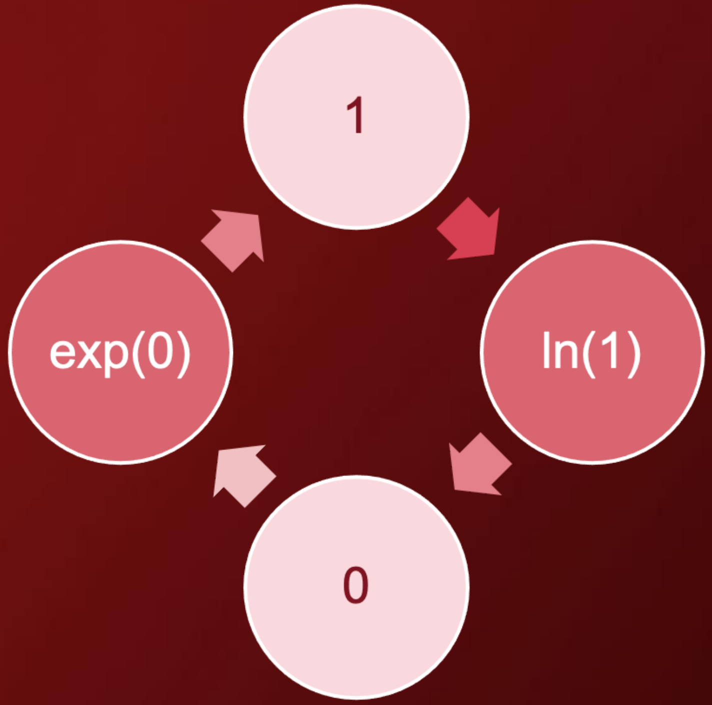
Outcome

- We predict the log odds of the outcome occurring

$\hat{b}_0$ and $\hat{b}_1$

- Note the the logistic regression equation is the same as the linear model, except we predict the log odds of the outcome
- $\hat{b}_1$ is the change in the log odds of the outcome associated with a unit change in the predictor

Logs and exponents



The odds ratio: exp(b)

$$\exp(b) = \frac{\text{odds after a unit change in the predictor}}{\text{original odds}}$$

b_0

- Log odds of outcome when the predictors are 0
- Easier to interpret exp(b_0), the odds of outcome when predictor is 0

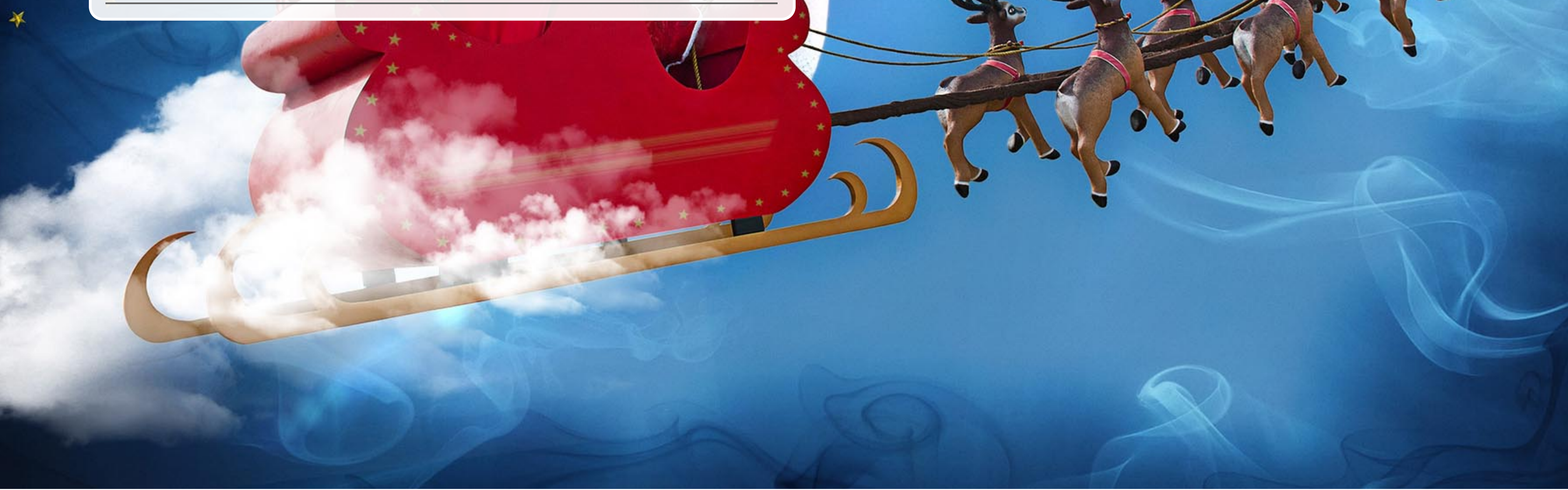
b_1

- Change in the log odds of outcome associated with a unit change in the predictor
- Easier to interpret exp (b_1), the odds ratio associated with a unit change in the predictor
- OR > 1: Predictor ↑, probability of outcome occurring ↑
- OR < 1: Predictor ↑, probability of outcome occurring ↓

Classification table

	Delivered	Not delivered	Total
Christmas pudding	150	28	178
Mulled wine	100	122	222
Total	250	150	400

$$\text{odds}_{\text{delivery}} = \frac{\text{Number delivered}}{\text{Number not delivered}} = \frac{250}{150} = 1.67$$



The odds ratio

$$\text{odds}_{\text{delivered after pudding}} = \frac{\text{Number delivered after pudding}}{\text{Number not delivered after pudding}}$$
$$= \frac{150}{28}$$
$$= 5.36$$

The odds ratio

$$
\begin{aligned}
\text{odds}_{\text{delivered after pudding}} &= \frac{\text{Number delivered after pudding}}{\text{Number not delivered after pudding}} \\
&= \frac{150}{28} \\
&= 5.36
\end{aligned}
$$

$$
\begin{aligned}
\text{odds}_{\text{delivered after wine}} &= \frac{\text{Number delivered after wine}}{\text{Number not delivered after wine}} \\
&= \frac{100}{122} \\
&= 0.82
\end{aligned}
$$

The odds ratio

$$\text{odds ratio} = \frac{\text{odds}_{\text{delivered after wine}}}{\text{odds}_{\text{delivered after pudding}}} = \frac{0.82}{5.36} = 0.15$$

The odds ratio

$$\begin{aligned}\text{odds ratio} &= \frac{\text{odds}_{\text{delivered after pudding}}}{\text{odds}_{\text{delivered after wine}}} \\ &= \frac{5.36}{0.82} \\ &= 6.54\end{aligned}$$

Fitting the model

```
santa_mod <- glm(delivered ~ treat, data = santa_tib, family = binomial())

santa_mod %>%
  parameters::parameters() %>%
  insight::parameters_table(p_digits = 3) #for nice formatting
```

Parameter	Coefficient	SE	95% CI	z	p
(Intercept)	1.68	0.21	[1.29, 2.10]	8.15	< .001
treatMulled wine	-1.88	0.25	[-2.37, -1.41]	-7.63	< .001

Odds ratios

```
santa_mod %>%
  parameters::parameters(exponentiate = TRUE) %>%
  insight::parameters_table(p_digits = 3) #for nice formatting
```

Parameter	Coefficient	SE	95% CI	z	p
(Intercept)	5.36	1.10	[3.64, 8.18]	8.15	< .001
treatMulled wine	0.15	0.04	[0.09, 0.24]	-7.63	< .001

A festive example

Santa Claus wanted to test the effects of different types of treats and the quantity of them consumed on whether presents got delivered:

- Predictors
  - treat: Christmas pudding, Mulled wine
  - quantity: 0, 1, 2, 3, or 4
- Outcome
  - delivered: Did the presents get delivered?

The data in R

Table 1: Santa's data

	id	quantity	treat	delivered
1	Fankle the Determined	4	Mulled wine	Not delivered
2	Amber the Cuddle	3	Mulled wine	Not delivered
3	Snorklum the Content	3	Pudding	Delivered
4	Henri the Biddible	0	Pudding	Delivered
5	Ainsley the Eigen vector	0	Mulled wine	Delivered
6	Funnelcup the Iron maiden fan	1	Mulled wine	Delivered
7	Ramsey the Surprised	2	Pudding	Delivered
8	Bindlestiff the Surprised	3	Mulled wine	Not delivered
9	Lardy the Strong	1	Pudding	Delivered
10	Theo the Small	1	Pudding	Delivered

Previous 1 2 3 4 5 … 40 Next

Extending the model

$$\text{delivered}_i = \hat{b}_0 + \hat{b}_1\text{treat}_i + \hat{b}_2\text{quantity}_i + \hat{b}_3\text{treat} \times \text{quantity}_i + e_i$$

Group	Treat
Christmas pudding	0
Mulled wine	1

Extending the model

$$\text{delivered}_i = \hat{b}_0 + \hat{b}_1\text{treat}_i + \hat{b}_2\text{quantity}_i + \hat{b}_3\text{treat} \times \text{quantity}_i + e_i$$

Group	Treat
Christmas pudding	0
Mulled wine	1

$$P(\text{delivery}) = \frac{1}{1 + e^{-(\hat{b}_0 + \hat{b}_1\text{treat}_i + \hat{b}_2\text{quantity}_i + \hat{b}_3\text{treat}\times\text{quantity}_i + e_i)}}$$

Extending the model

$$\text{delivered}_i = \hat{b}_0 + \hat{b}_1\text{treat}_i + \hat{b}_2\text{quantity}_i + \hat{b}_3\text{treat} \times \text{quantity}_i + e_i$$

Group	Treat
Christmas pudding	0
Mulled wine	1

$$P(\text{delivery}) = \frac{1}{1 + e^{-(\hat{b}_0 + \hat{b}_1\text{treat}_i + \hat{b}_2\text{quantity}_i + \hat{b}_3\text{treat}\times\text{quantity}_i + e_i)}}$$

$$\ln\left(\frac{P(\text{delivery})}{1 - P(\text{delivery})}\right) = \hat{b}_0 + \hat{b}_1\text{treat}_i + \hat{b}_2\text{quantity}_i + \hat{b}_3\text{treat} \times \text{quantity}_i + e_i$$

Building the model

- Forced entry: all variables entered simultaneously.
- Hierarchical: variables entered in blocks.
    - Blocks should be based on past research, or theory being tested.
    - Good method.
- Stepwise: variables entered on the basis of statistical criteria (i.e., Relative contribution to predicting outcome).
    - Should be used only for exploratory analysis.

Things that can go wrong

Things *we've* met before

- Linearity (of the logit)
- Spherical residuals
    - Independent errors
- Multicollinearity

Unique problems

- Incomplete information
- Complete separation

Incomplete information

Empty cells

- We don't know how many presents are delivered after two puddings or not delivered after 5 wines
- Problem quickly escalates with continuous predictors
- Inflates standard errors

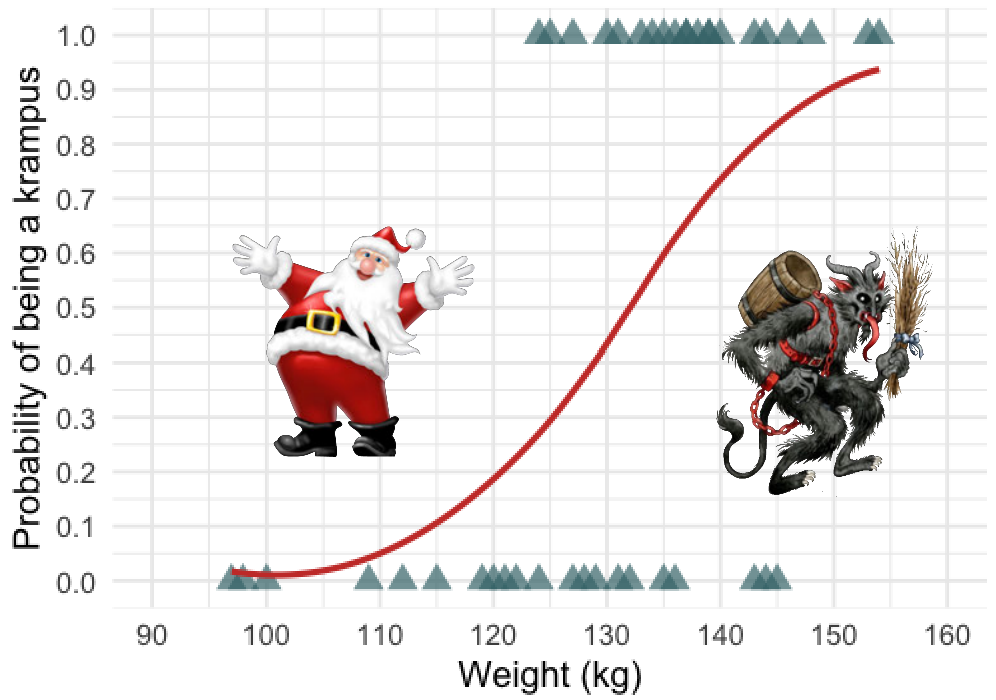
	Quantity	Delivered	Not delivered
Pudding	0	27	3
Pudding	1	36	10
Pudding	2	-	5
Pudding	3	37	6
Pudding	4	12	4
Mulled wine	0	27	3
Mulled wine	1	32	11
Mulled wine	2	24	36
Mulled wine	3	14	48
Mulled wine	4	3	-

Complete separation

When the outcome variable can be perfectly predicted

- Predicting whether someone is a krampus based on weight

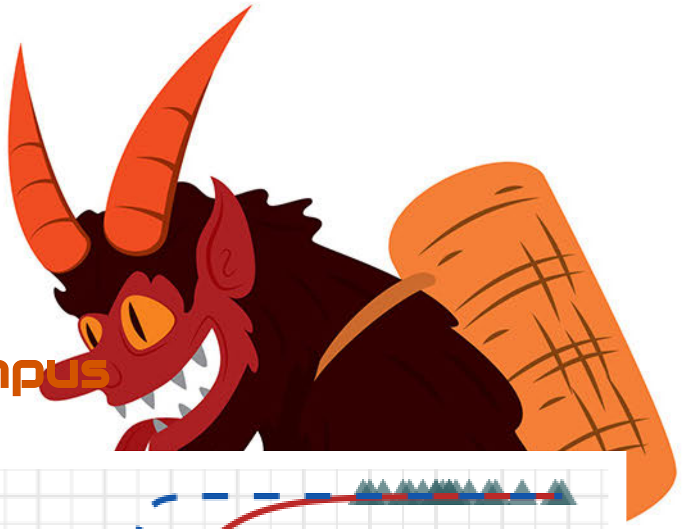
Santa vs. krampus



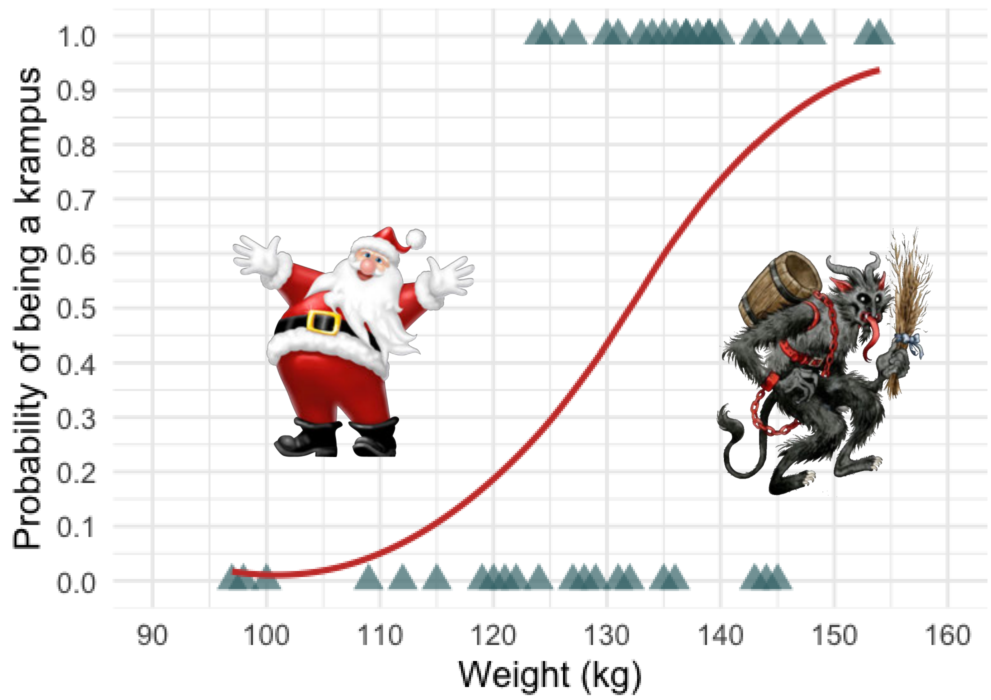
Complete separation

When the outcome variable can be perfectly predicted

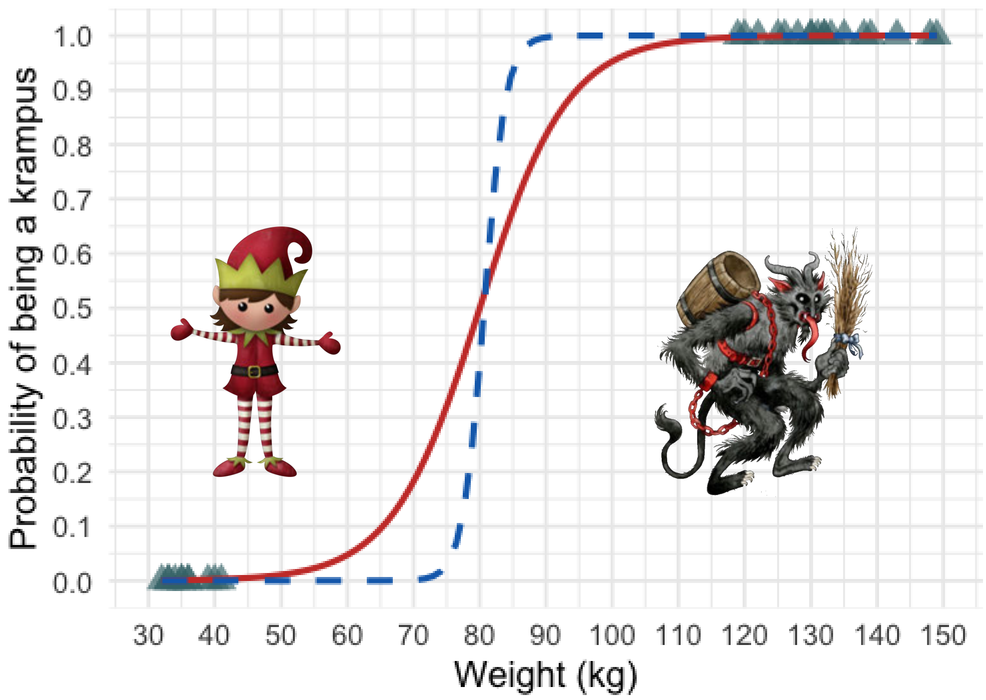
- Predicting whether someone is a krampus based on weight



Santa vs. krampus



Elves vs. krampus



Fitting the model

```
santa_full_mod <- glm(delivered ~ treat*quantity, data = santa_tib, family = binomial())

santa_full_mod %>%
  parameters::parameters() %>%
  insight::parameters_table(p_digits = 3) #for nice formatting
```

Fitting the model

```
santa_full_mod <- glm(delivered ~ treat*quantity, data = santa_tib, family = binomial())

santa_full_mod %>%
  parameters::parameters() %>%
  insight::parameters_table(p_digits = 3) #for nice formatting
```

Parameter	Coefficient	SE	95% CI	z	p
(Intercept)	1.83	0.38	[1.13, 2.63]	4.80	< .001
treatMulled wine	0.20	0.52	[-0.83, 1.22]	0.38	0.701
quantity	-0.08	0.17	[-0.41, 0.25]	-0.48	0.629
treatMulled wine:quantity	-1.03	0.23	[-1.49, -0.58]	-4.45	< .001

```
santa_tib %>%
  filter(treat == "Pudding") %>%
  glm(delivered ~ quantity, data = ., family
= binomial())
```

Parameter	Coefficient	SE	95% CI	z	p
(Intercept)	1.83	0.38	[1.13, 2.63]	4.80	< .001
quantity	-0.08	0.17	[-0.41, 0.25]	-0.48	0.629

```
santa_tib %>%
  filter(treat == "Pudding") %>%
  glm(delivered ~ quantity, data = ., family
= binomial())
```

Parameter	Coefficient	SE	95% CI	z	p
(Intercept)	1.83	0.38	[1.13, 2.63]	4.80	< .001
quantity	-0.08	0.17	[-0.41, 0.25]	-0.48	0.629

```
santa_tib %>%
  filter(treat == "Mulled wine") %>%
  glm(delivered ~ quantity, data = ., family
= binomial())
```

Parameter	Coefficient	SE	95% CI	z	p
(Intercept)	2.03	0.35	[1.37, 2.76]	5.73	< .001
quantity	-1.11	0.16	[-1.44, -0.81]	-6.99	< .001

```r
santa_tib %>%
  filter(treat == "Pudding") %>%
  glm(delivered ~ quantity, data = ., family
= binomial())
```

Parameter	Coefficient	SE	95% CI	z	p
(Intercept)	1.83	0.38	[1.13, 2.63]	4.80	< .001
quantity	-0.08	0.17	[-0.41, 0.25]	-0.48	0.629

```r
santa_tib %>%
  filter(treat == "Mulled wine") %>%
  glm(delivered ~ quantity, data = ., family
= binomial())
```

Parameter	Coefficient	SE	95% CI	z	p
(Intercept)	2.03	0.35	[1.37, 2.76]	5.73	< .001
quantity	-1.11	0.16	[-1.44, -0.81]	-6.99	< .001

$$
\begin{aligned}
b_{\text{wine}} - b_{\text{pudding}} &= -1.11 - (-0.08) \\
&= -1.03
\end{aligned}
$$

Odds ratio

```
santa_full_mod %>%
  parameters::parameters(exponentiate = TRUE) %>%
  insight::parameters_table(p_digits = 3) #for nice formatting
```

Parameter	Coefficient	SE	95% CI	z	p
(Intercept)	6.23	2.37	[3.08, 13.86]	4.80	< .001
treatMulled wine	1.22	0.63	[0.43, 3.37]	0.38	0.701
quantity	0.92	0.15	[0.66, 1.28]	-0.48	0.629
treatMulled wine:quantity	0.36	0.08	[0.23, 0.56]	-4.45	< .001

```
santa_tib %>%
  filter(treat == "Pudding") %>%
  glm(delivered ~ quantity, data = ., family
= binomial())
```

Parameter	Coefficient	SE	95% CI	z	p
(Intercept)	1.83	0.38	[1.13, 2.63]	4.80	< .001
quantity	-0.08	0.17	[-0.41, 0.25]	-0.48	0.629

```
santa_tib %>%
  filter(treat == "Mulled wine") %>%
  glm(delivered ~ quantity, data = ., family
= binomial())
```

Parameter	Coefficient	SE	95% CI	z	p
(Intercept)	2.03	0.35	[1.37, 2.76]	5.73	< .001
quantity	-1.11	0.16	[-1.44, -0.81]	-6.99	< .001

$$
\begin{aligned}
b_{\text{wine}} - b_{\text{pudding}} &= -1.11 - (-0.08) \\
&= -1.03
\end{aligned}
$$

```r
santa_tib %>%
  filter(treat == "Pudding") %>%
  glm(delivered ~ quantity, data = ., family
= binomial())
```

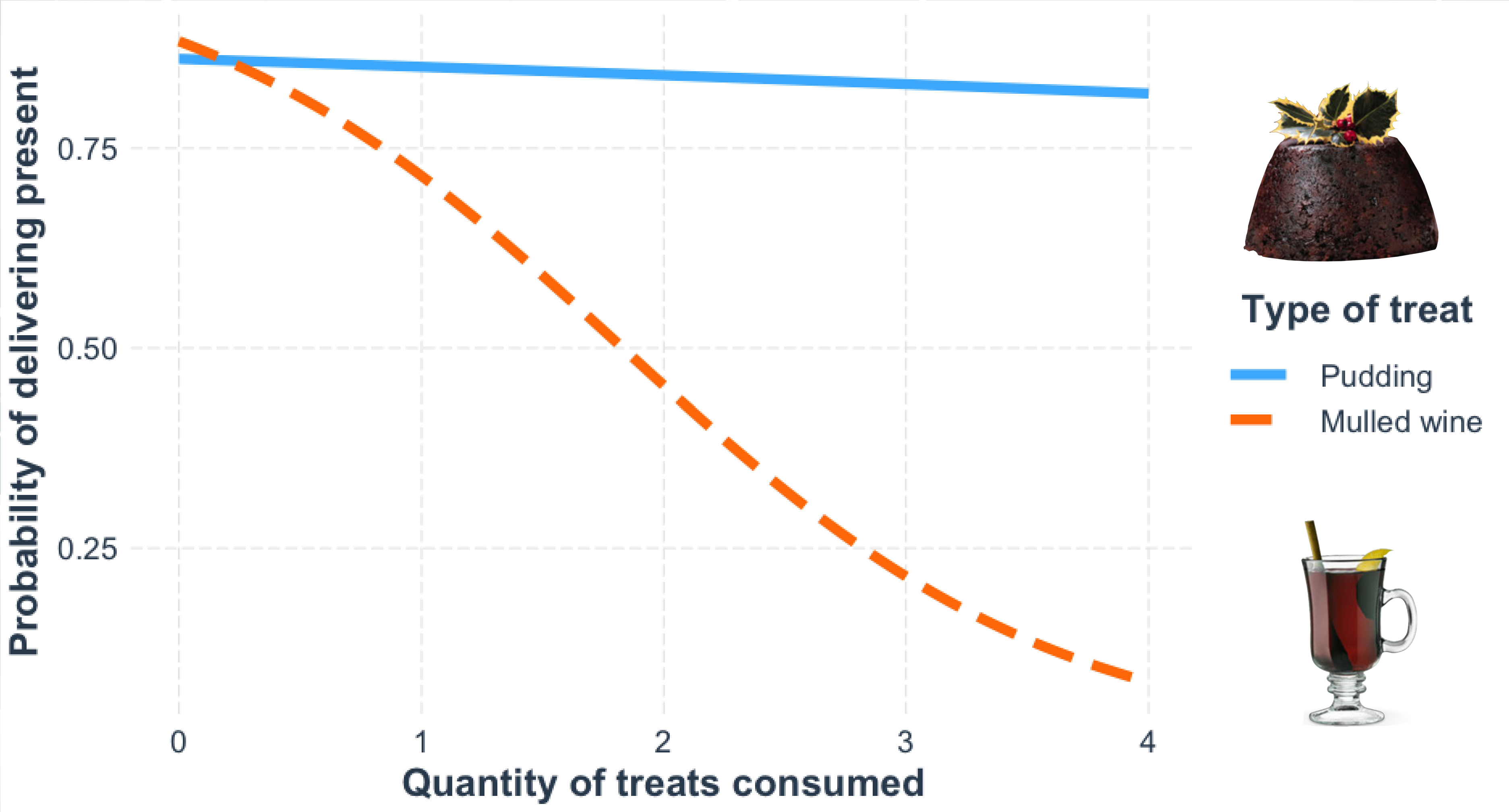
Parameter	Coefficient	SE	95% CI	z	p
(Intercept)	1.83	0.38	[1.13, 2.63]	4.80	< .001
quantity	-0.08	0.17	[-0.41, 0.25]	-0.48	0.629

```r
santa_tib %>%
  filter(treat == "Mulled wine") %>%
  glm(delivered ~ quantity, data = ., family
= binomial())
```

Parameter	Coefficient	SE	95% CI	z	p
(Intercept)	2.03	0.35	[1.37, 2.76]	5.73	< .001
quantity	-1.11	0.16	[-1.44, -0.81]	-6.99	< .001

$$
\begin{aligned}
b_{\text{wine}} - b_{\text{pudding}} &= -1.11 - (-0.08) \\
&= -1.03
\end{aligned}
$$

$$
\begin{aligned}
\exp(b)_{\text{difference}} &= e^{-1.03} \\
&= 0.36
\end{aligned}
$$


Probability of delivering present
0.75
0.50
0.25
0
1
2
3
4
Quantity of treats consumed
Type of treat
Pudding
Mulled wine