Logistic Regression Introduction

Julie Deeke
Statistics with Python Course Developer

Cartwheel Data

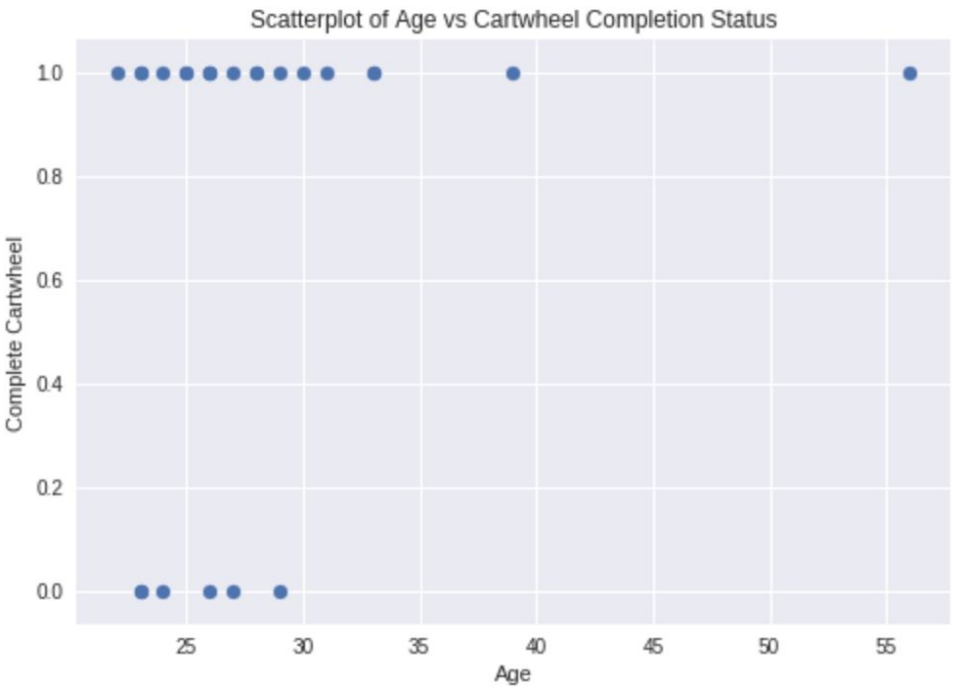
Random sample of 25 adults attempted cartwheels

Primary Variable of interest: Cartwheel completion

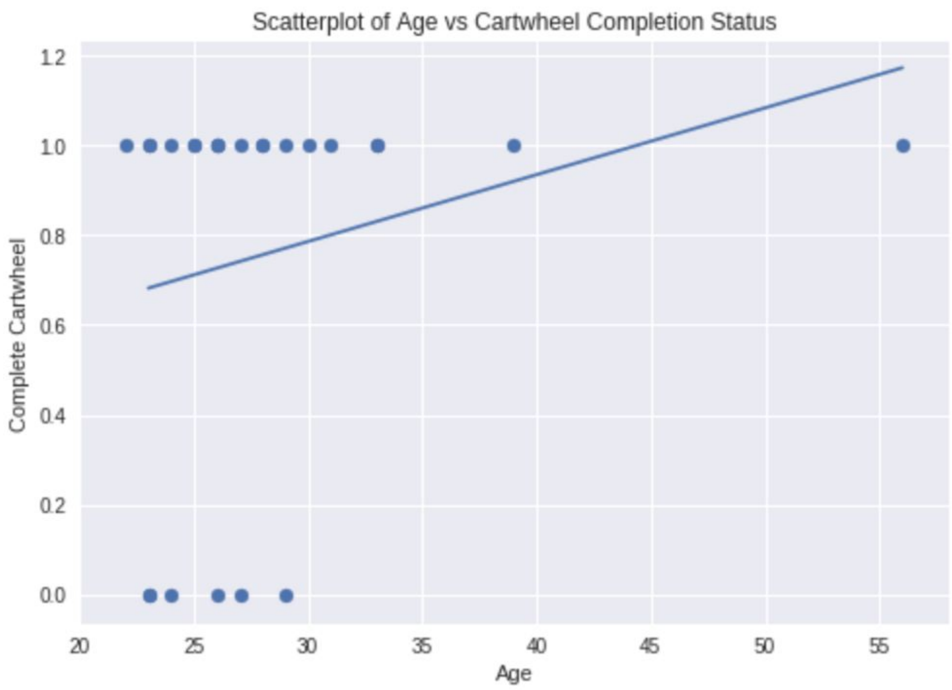
Research Question

Based on age, can we predict whether a cartwheel is completed?

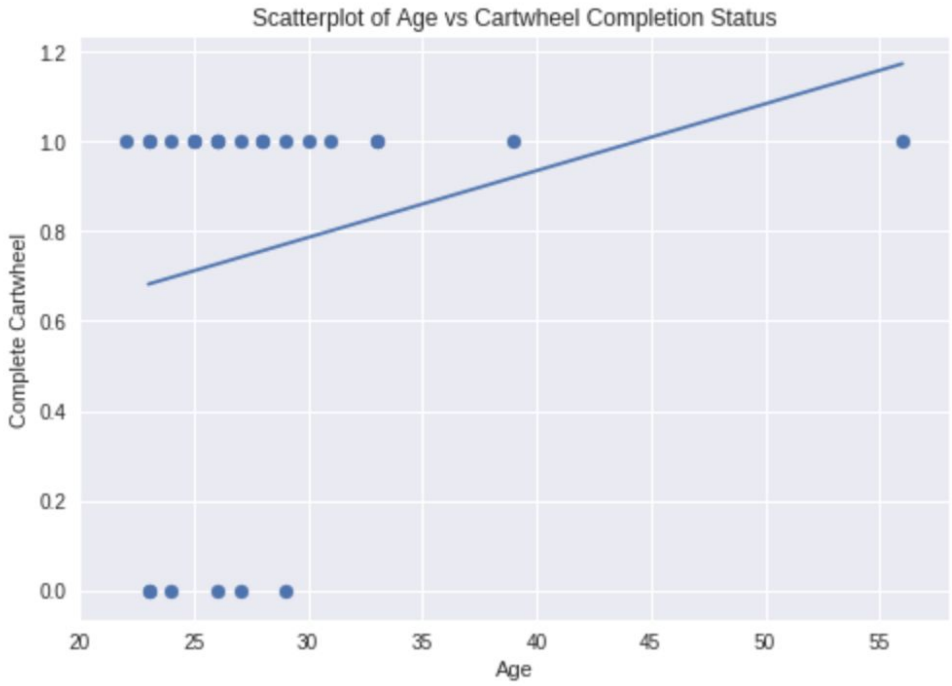
Let's Look at the Data



Linear Model



Linear Model



$\hat{y} = 0.34 + 0.015\ \text{age}$

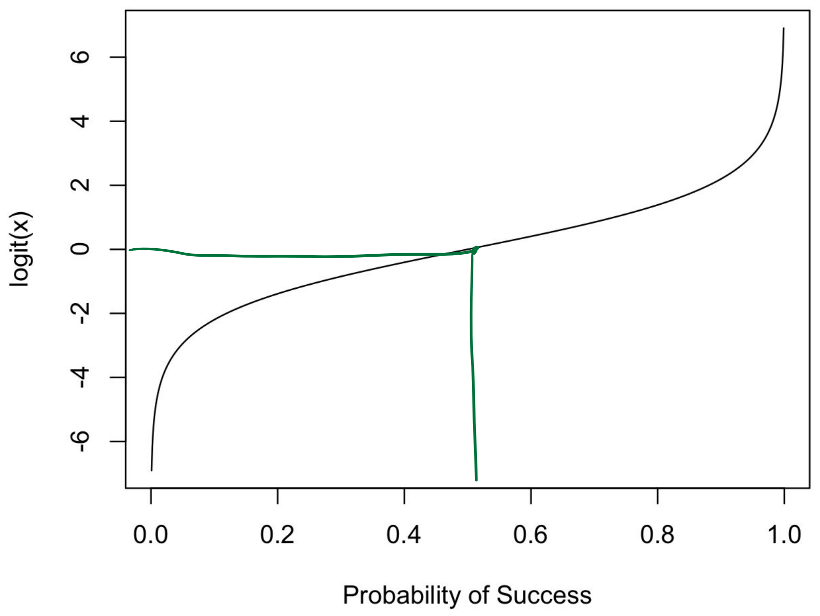
Logit Transformation

- Instead of predicting completion status, we predict a ***transformed version*** of the probability of a success

Logit Transformation

- Instead of predicting completion status, we predict a ***transformed version*** of the probability of a success
- Uses the logit function $\ln(\frac{p}{1-p})$

symmetric

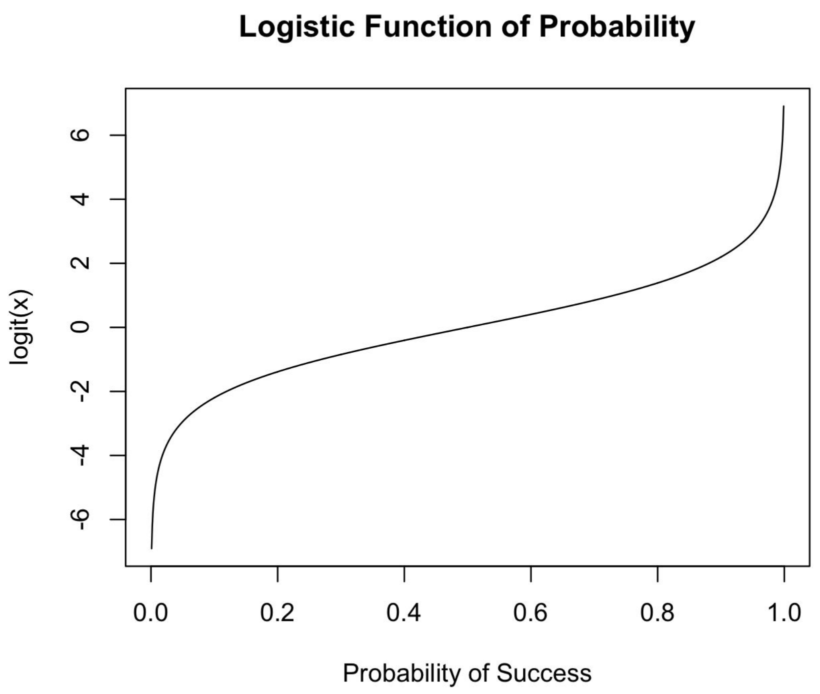


Logit Transformation

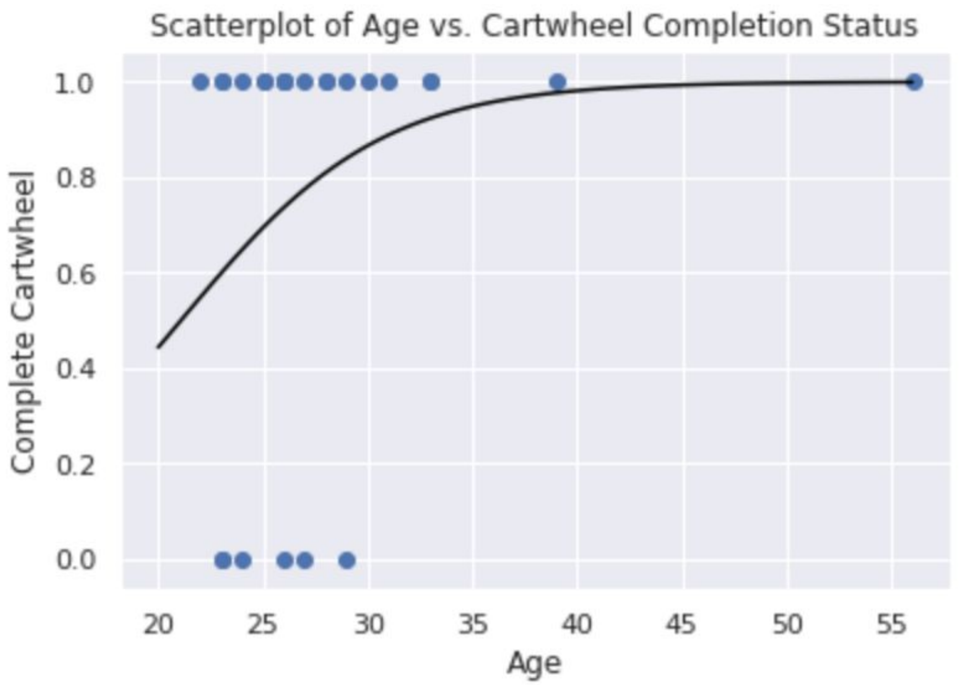
- Instead of predicting completion status, we predict a ***transformed version*** of the probability of a success
- Uses the logit function:

$\ln\left(\frac{p}{1-p}\right)$

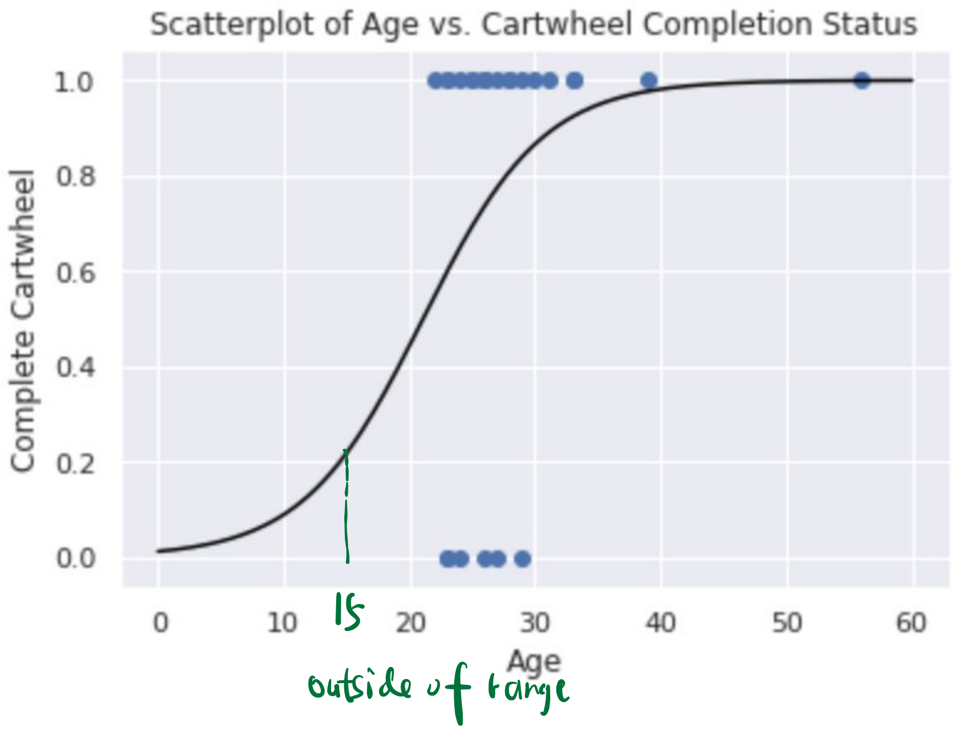
- $\text{logit}(\hat{y}) = b_0 + b_1x$



Logistic Regression Line



Logistic Regression Line



Extrapolation IVQ

Would you feel comfortable using this model to estimate the probability that a teenager who is 15 can complete a cartwheel?

Logistic Regression Equation

Generalized Linear Model Regression Results

Dep. Variable:	CompleteGroup	**No. Observations:**	25
Model:	GLM	**Df Residuals:**	23
Model Family:	Binomial	**Df Model:**	1
Link Function:	logit	**Scale:**	1.0

	coef	**std err**	**z**	**P>|z|**	**[0.025**	**0.975]**
Intercept	-4.4213	4.429	-0.998	0.318	-13.101	4.259
Age	0.2096	0.171	1.225	0.221	-0.126	0.545

Logistic Regression Equation

Generalized Linear Model Regression Results

Dep. Variable:	CompleteGroup	**No. Observations:**	25
Model:	GLM	**Df Residuals:**	23
Model Family:	Binomial	**Df Model:**	1
Link Function:	logit	**Scale:**	1.0

	coef	**std err**	**z**	**P>|z|**	**[0.025**	**0.975]**
Intercept	-4.4213	4.429	-0.998	0.318	-13.101	4.259
Age	0.2096	0.171	1.225	0.221	-0.126	0.545

Logistic Regression Equation

$$\text{logit}(\ \hat{y}\ ) = -4.42 + 0.2096 \text{ age}$$

Generalized Linear Model Regression Results

Dep. Variable:	CompleteGroup	**No. Observations:**	25
Model:	GLM	**Df Residuals:**	23
Model Family:	Binomial	**Df Model:**	1
Link Function:	logit	**Scale:**	1.0

	coef	std err	z	P>|z|	[0.025	0.975]
Intercept	-4.4213	4.429	-0.998	0.318	-13.101	4.259
Age	0.2096	0.171	1.225	0.221	-0.126	0.545

Logistic Regression Equation

$$\text{logit}(\ \hat{y}\ ) = -4.42 + 0.2096\ \text{age}$$

Slope interpretation:
For each increase in age by 1 year, the log odds of a successful cartwheel increases by about 0.2096, on average.

Generalized Linear Model Regression Results

Dep. Variable:	CompleteGroup	**No. Observations:**	25
Model:	GLM	**Df Residuals:**	23
Model Family:	Binomial	**Df Model:**	1
Link Function:	logit	**Scale:**	1.0

	coef	**std err**	**z**	**P>|z|**	**[0.025**	**0.975]**
Intercept	-4.4213	4.429	-0.998	0.318	-13.101	4.259
Age	0.2096	0.171	1.225	0.221	-0.126	0.545

Logistic Regression Equation

$$\text{logit}(\hat{y}) = -4.42 + 0.2096\ \text{age}$$

Slope interpretation: For each year increase in age, the odds of a successful cartwheel increases by about 1.23 ($e^{0.2096}$) times that of the younger age, on average.

Generalized Linear Model Regression Results

Dep. Variable:	CompleteGroup	**No. Observations:**	25
Model:	GLM	**Df Residuals:**	23
Model Family:	Binomial	**Df Model:**	1
Link Function:	logit	**Scale:**	1.0

	coef	**std err**	**z**	**P>|z|**	**[0.025**	**0.975]**
Intercept	-4.4213	4.429	-0.998	0.318	-13.101	4.259
Age	0.2096	0.171	1.225	0.221	-0.126	0.545

Predicted Probability of Success

- For someone who is 36, what is their predicted log odds of a successful cartwheel using the model?

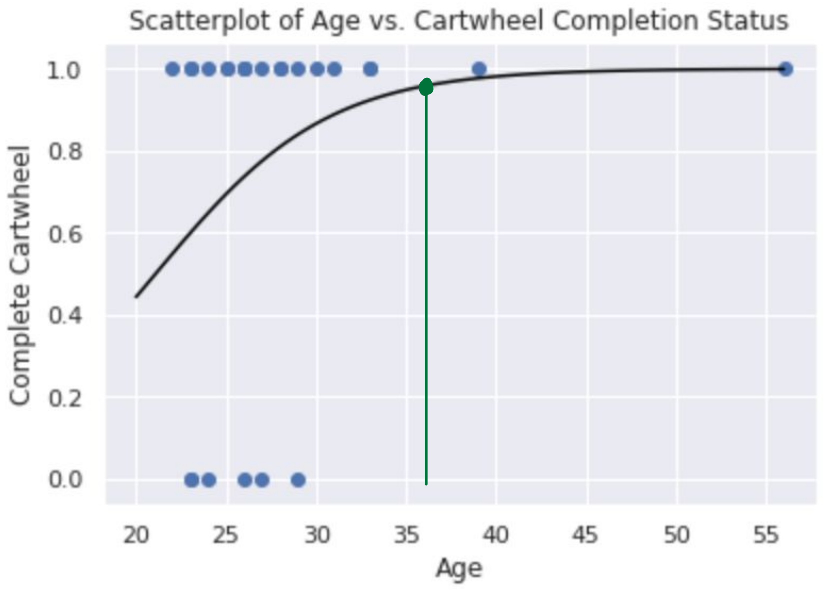
Predicted Probability of Success

- For someone who is 36, what is their predicted log odds of a successful cartwheel using the model?

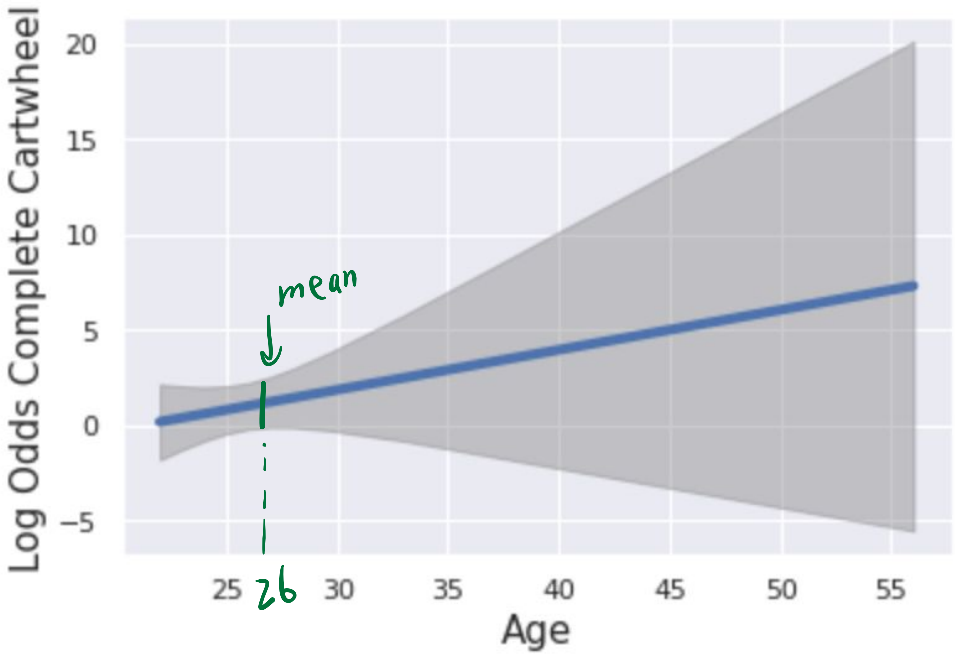
$$\begin{aligned}\text{logit}(\hat{y}) &= -4.42 + 0.2096\ \text{age} \\ &= -4.42 + 0.2096\ (36) \\ &= 3.13\end{aligned}$$

Predicted Probability of Success

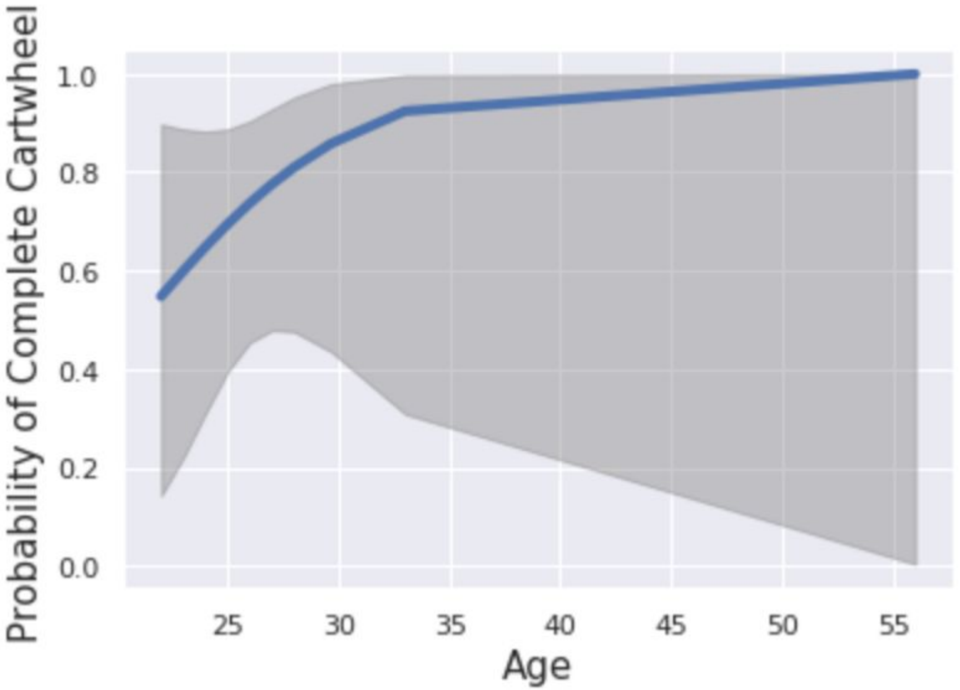
- For someone who is 36, what is their predicted log odds of a successful cartwheel using the model?
- Using the graph on the right, estimate what the probability of success might be?



Prediction Uncertainty



Prediction Uncertainty



Assumptions

We need to assume that our model $\text{logit}(y) = \beta_0 + \beta_1 x_1$ is appropriate

Assumptions

We need to assume that our model $\text{logit}(y) = \beta_0 + \beta_1 x_1$ is appropriate

~with a large enough sample size, you can identify discrepancies with residual plots

Assumptions

We need to assume that our model $\text{logit}(y) = \beta_0 + \beta_1 x_1$ is appropriate

~with a large enough sample size, you can identify discrepancies with residual plots

~y only takes two values, so residuals can be limited

Assumptions

We need to assume that our model $\text{logit}(y) = \beta_0 + \beta_1 x_1$ is appropriate

~with a large enough sample size, you can identify discrepancies with residual plots

~y only takes two values, so residuals can be limited

~to create informative residual plots, it helps if x takes a wide range of values and to have additional covariates