Multivariate: Logistic regression

Table of contents

1 Goals

1.1 Goals

1.1.1 Goals of this lecture

- My outcome variable **isn't normally distributed**
  - It's **binary**!!!
  - **Two mutually exclusive categories**
    - yes/no, pass/fail, diagnosed/not, etc.
  - Linear regression assumptions are violated
- Use **logistic regression** to analyze the outcome
  - It's an **extension** of linear regression, so many of the **same concepts** still apply

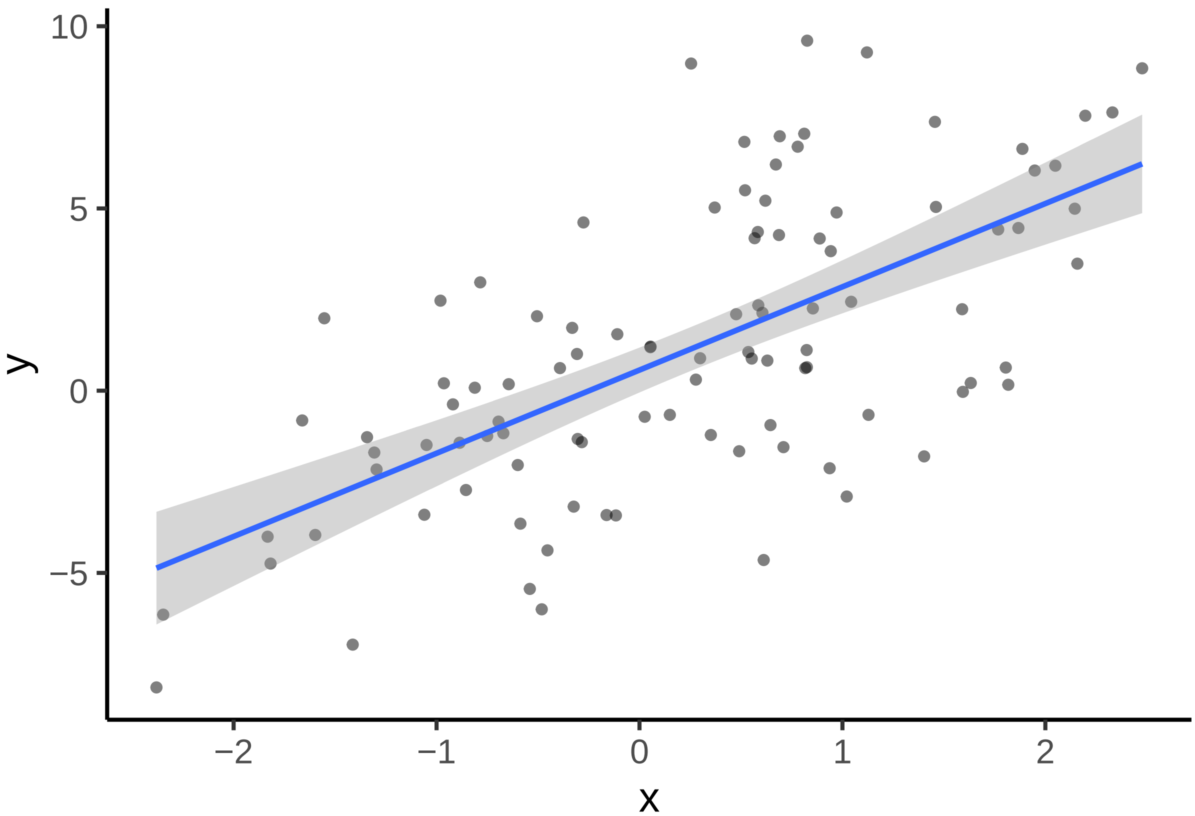
2 Linear regression and extensions

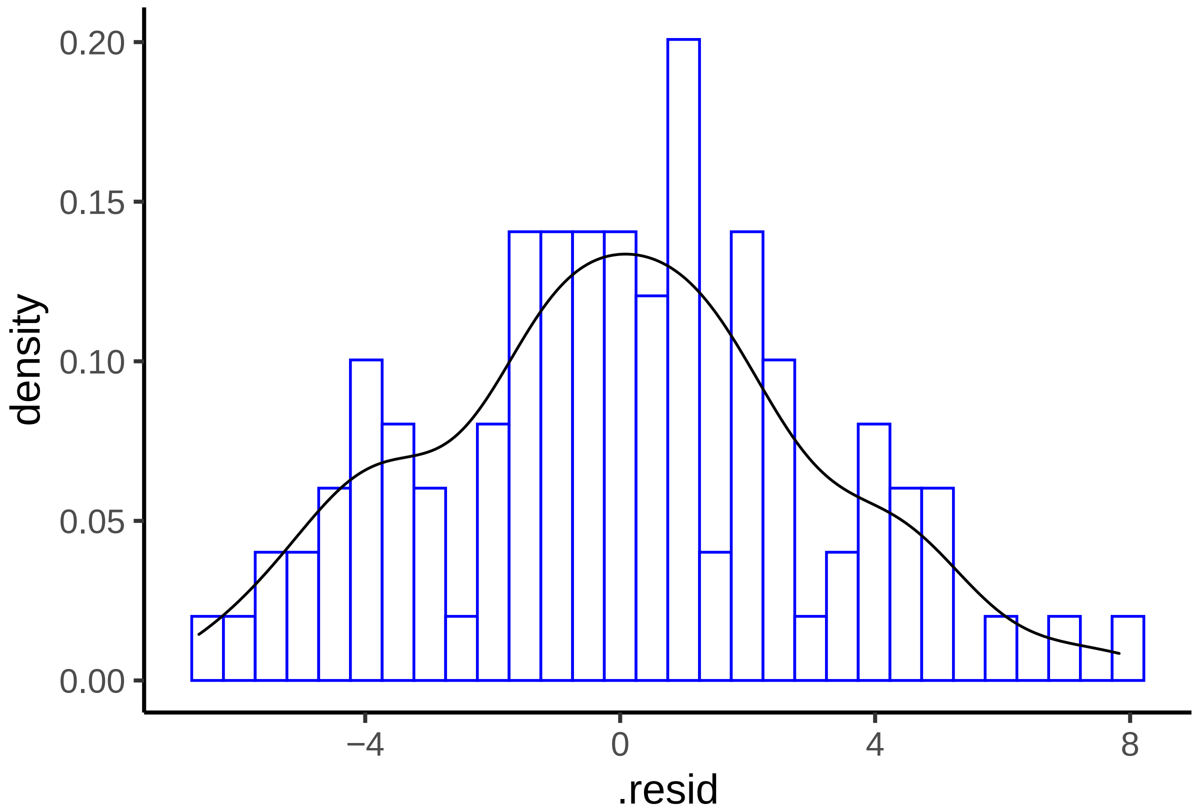
2.1 Review: Linear regression

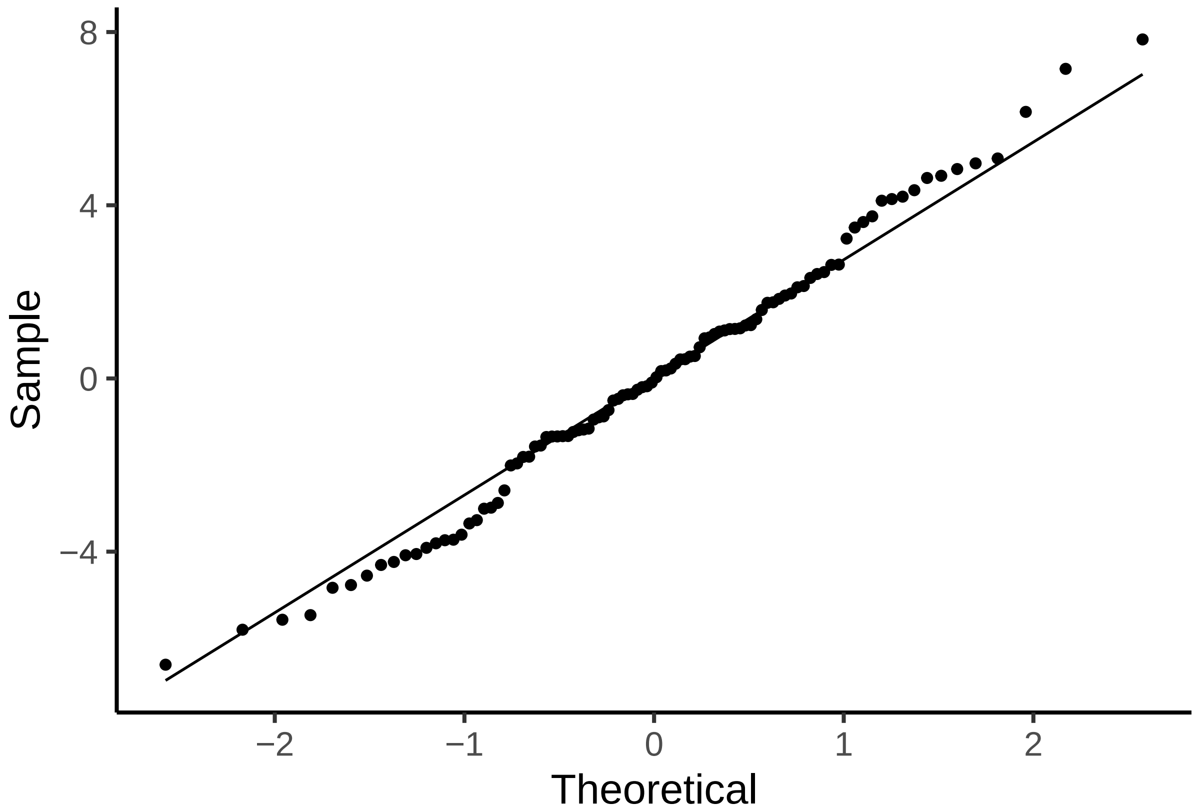
2.1.1 Assumptions of linear regression

General linear model (GLM, linear regression, ANOVA) makes **three assumptions** about the **residuals** ($e_i = Y_i - \hat{Y}_i$) of the model

1. **Independence**: observations (i.e., residuals) from different subjects **do not depend on one another**
2. **Constant variance** (homoscedasticity): **variance of residuals is same** at all values of predictor(s)
3. **Conditional normality**: **residuals are normally distributed** at each value of predictor(s)


10
5
0
−5
y
−2
−1
0
1
2
x


0.20
0.15
0.10
0.05
0.00
density
−4
0
4
8
.resid



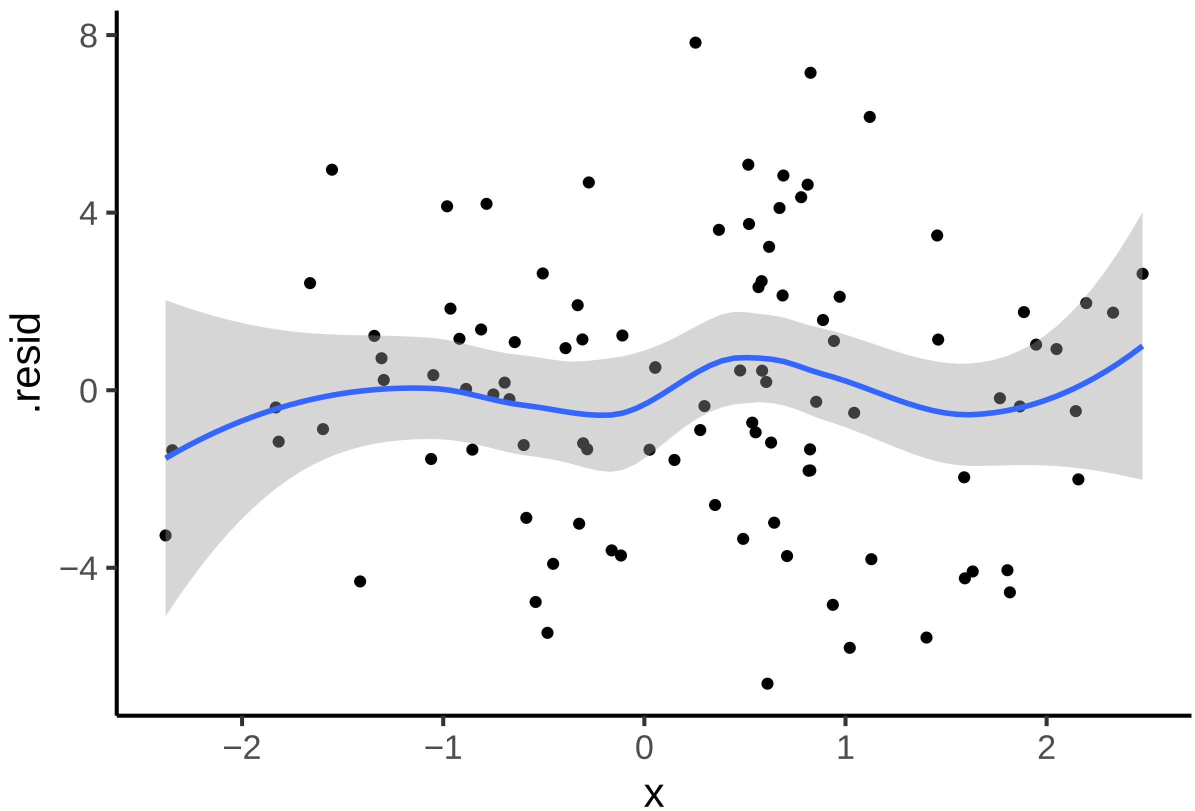
Sample
Theoretical
8
4
0
−4
−2
−1
0
1
2

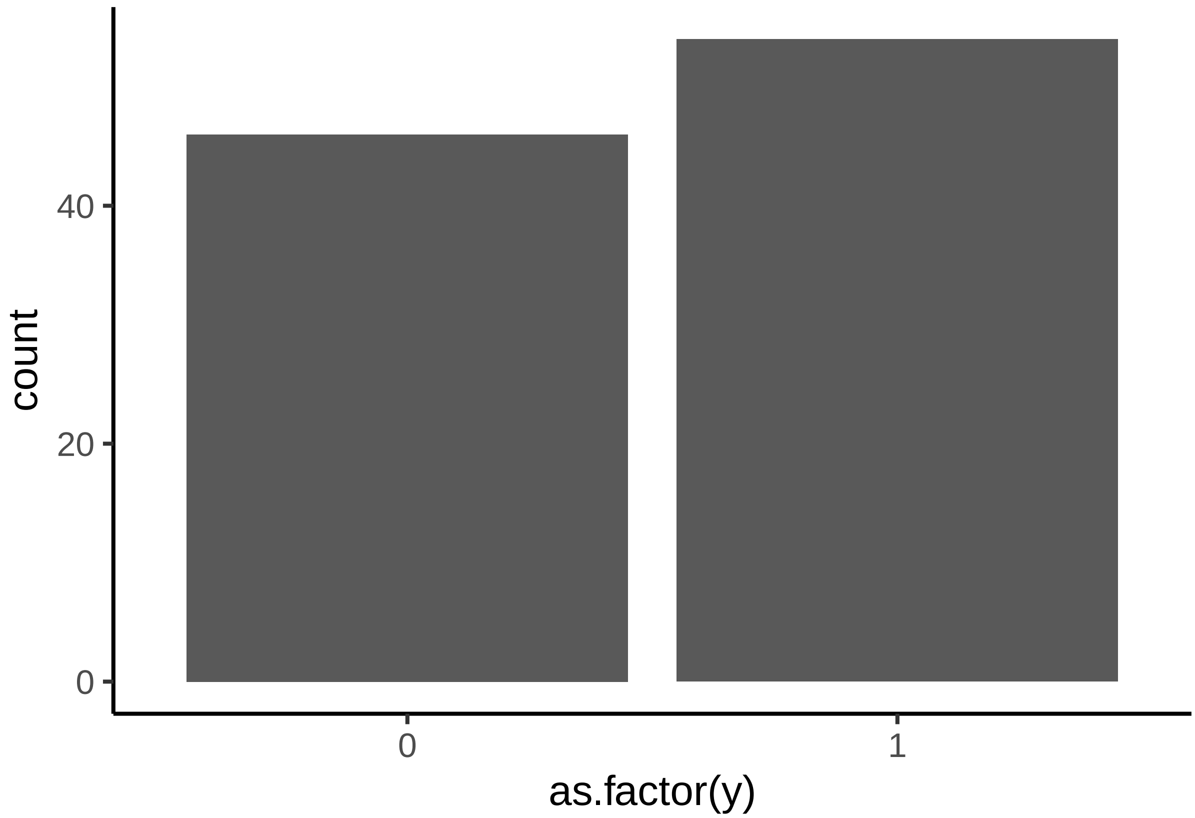
2.1.2 Linear regression on normal outcome

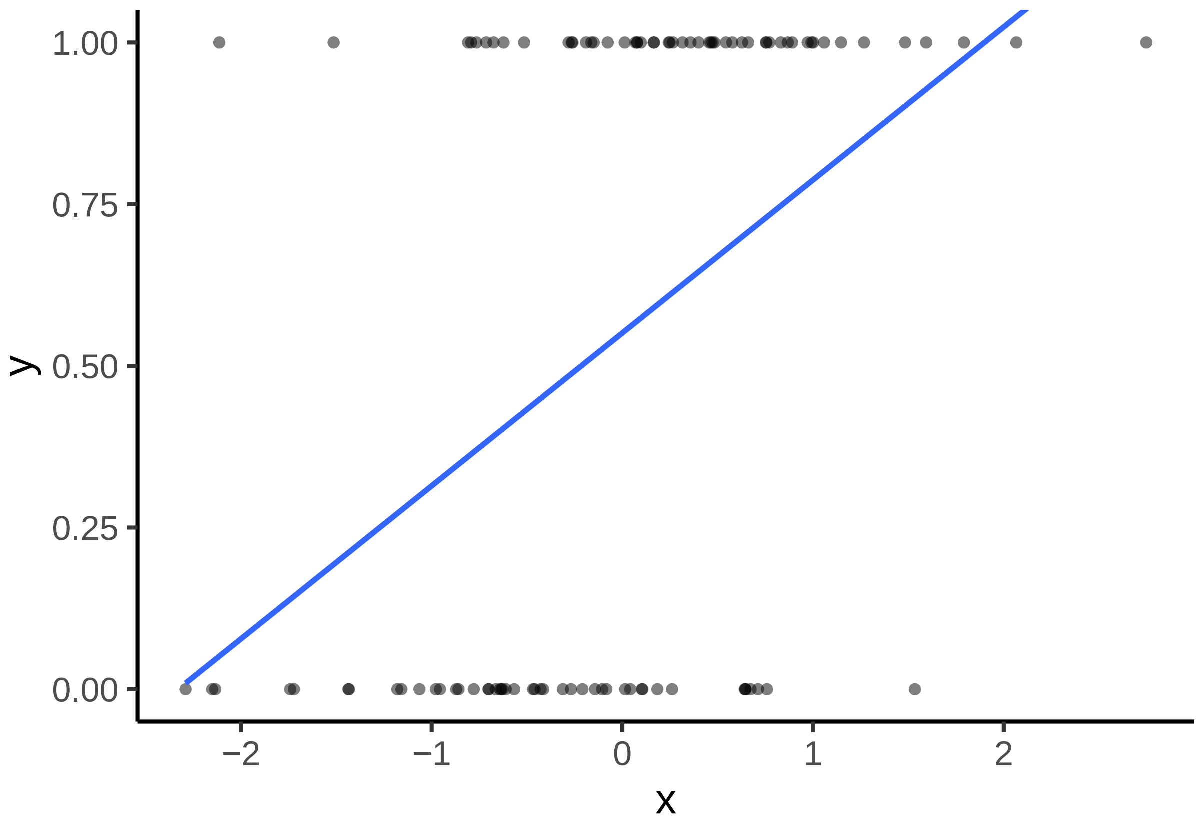
2.1.3 Assumptions met!

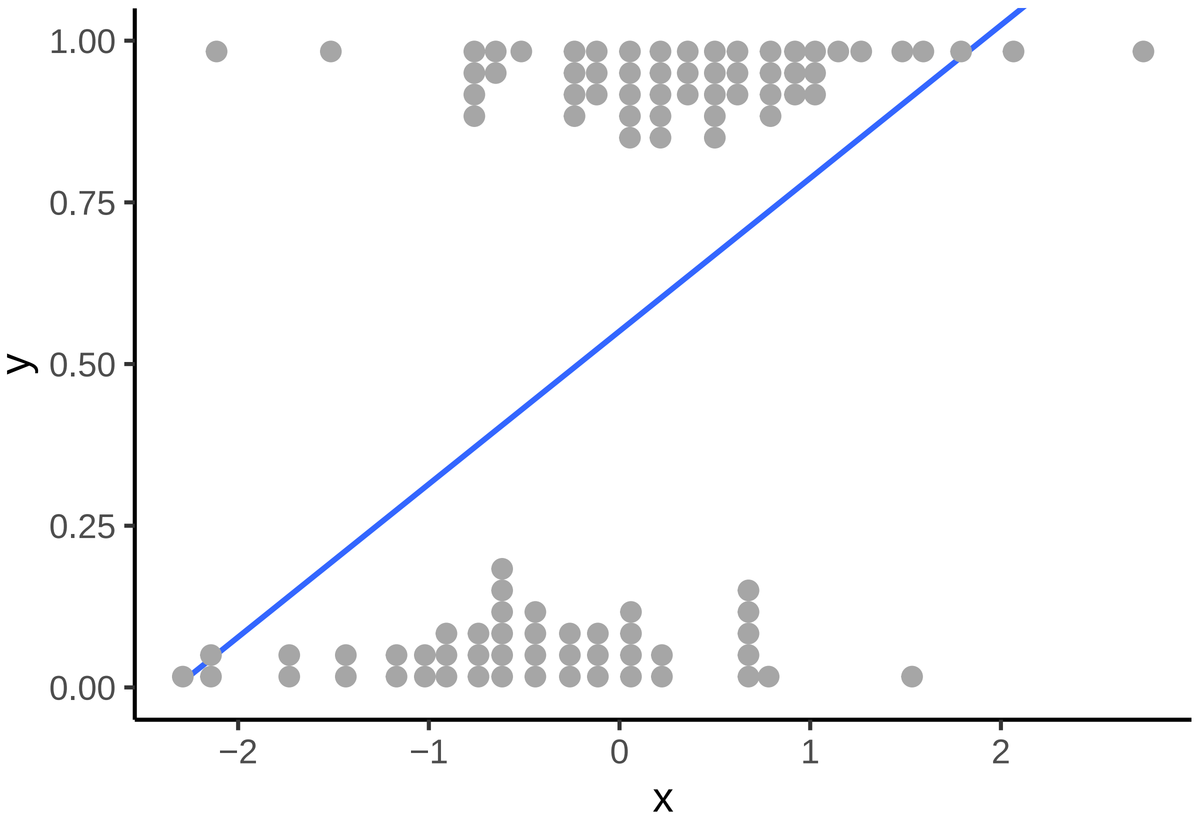
2.1.4 Assumptions met!

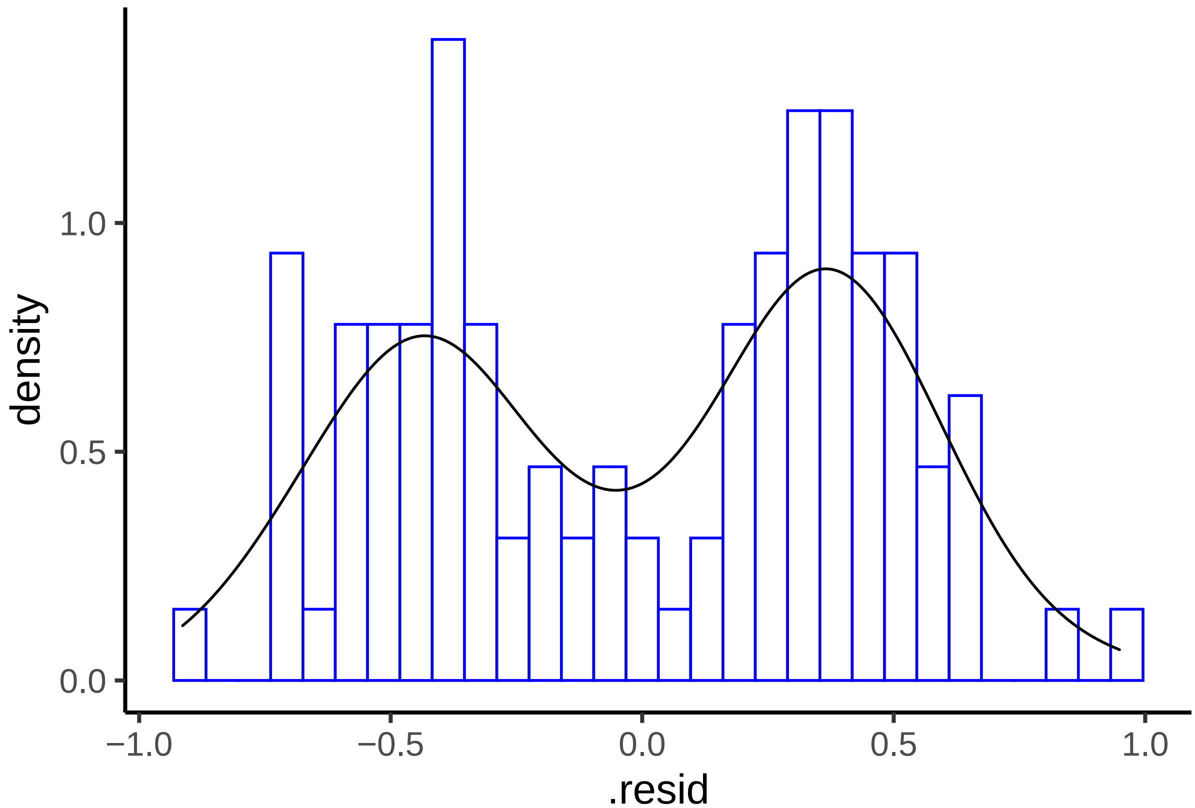
2.1.5 Assumptions met!




count
40
20
0
0
1
as.factor(y)


1.00
0.75
0.50
0.25
0.00
y
−2
−1
0
1
2
x


1.00
0.75
0.50
0.25
0.00
y
−2
−1
0
1
2
x


density
1.0
0.5
0.0
−1.0
−0.5
0.0
0.5
1.0
.resid

2.2 Linear regression with a binary variable

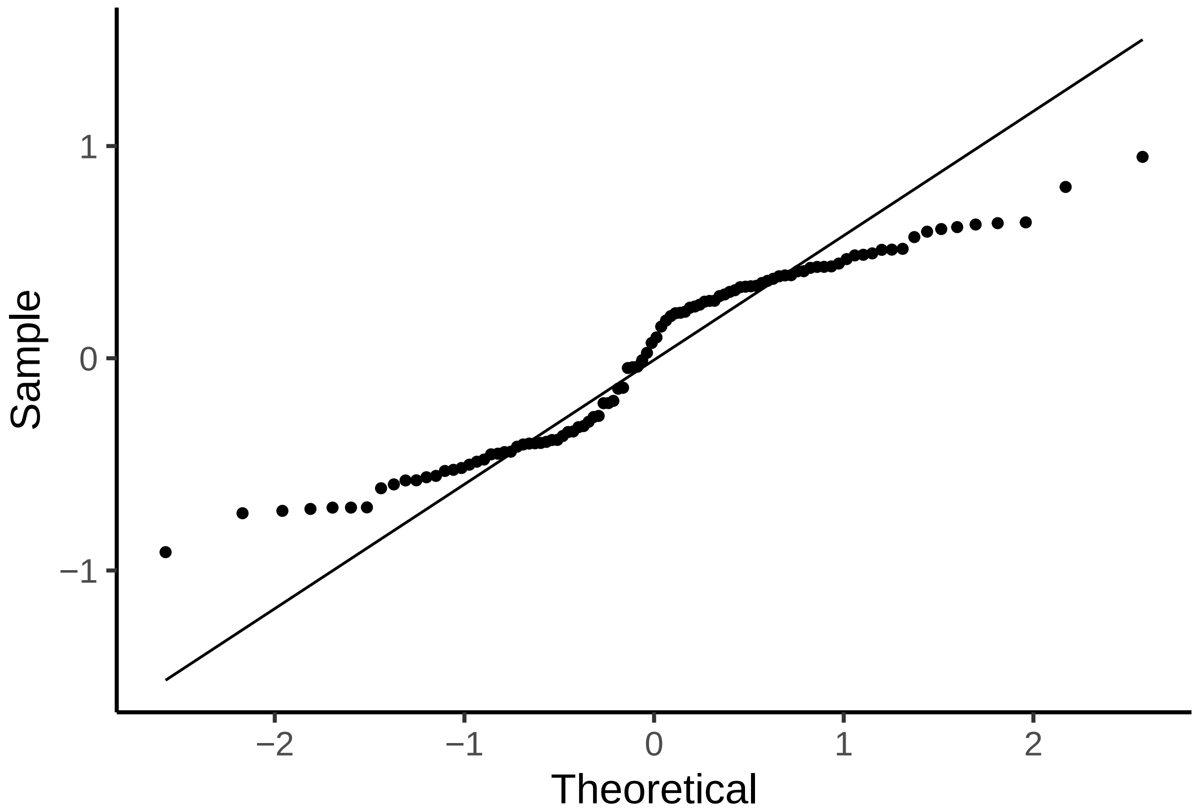
2.2.1 A binary variable is not normal

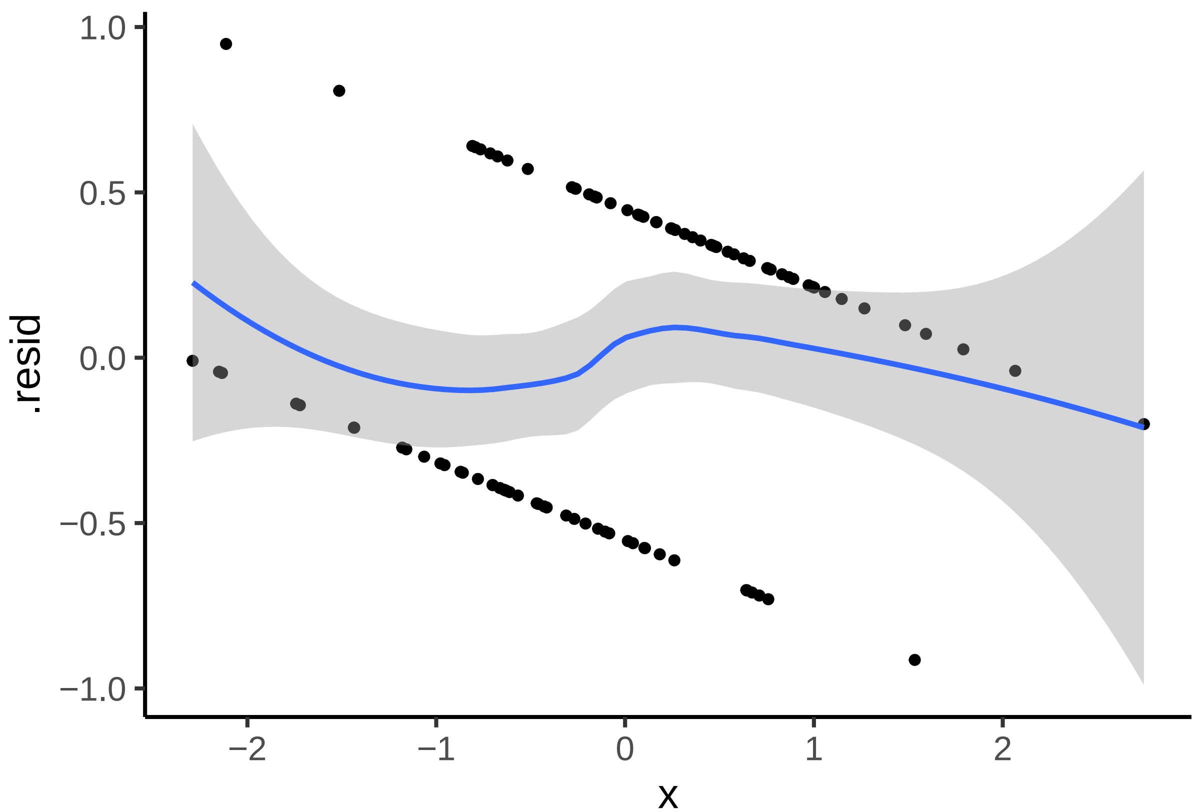
2.2.2 Plot of data with fit line

2.2.3 Plot of data with fit line

2.2.4 Plot of residuals

2.2.5 Plot of residuals





1.0
0.5
0.0
−0.5
−1.0
.resid
−2
−1
0
1
2
x

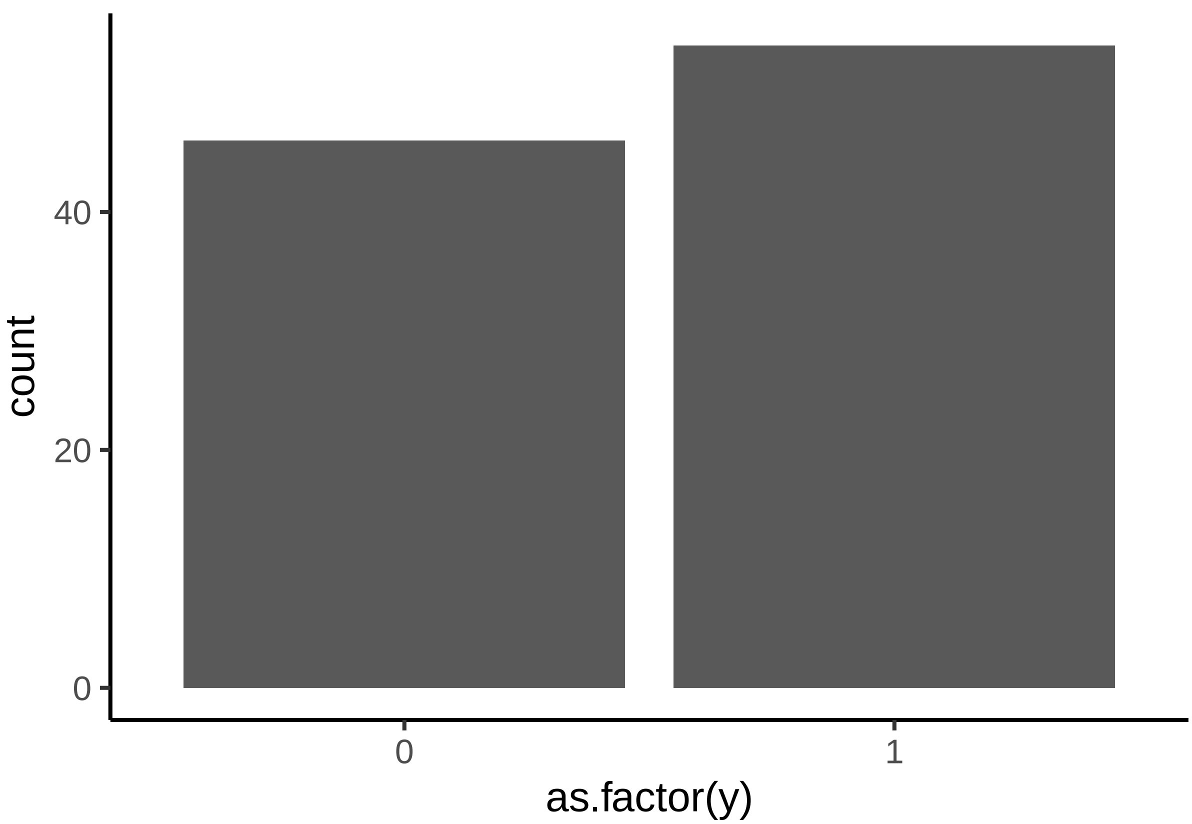
2.2.6 Plot of residuals

2.3 Next steps

2.3.1 What NOT to do

- **Ignore** the problem
  - Do linear regression anyway
  - Call it **linear probability model**
- **Transform** the outcome
  - Square root, natural log, etc.
  - May *slightly* normalize *univariate* residual distribution
  - **Does not fix heteroscedasticity, (conditional) non-normality**

2.3.2 A binary variable is not normal



2.3.3 What to do

The **generalized linear model (GLiM)**

- Not a single model but a **family** of regression models
- **Choose** features (e.g., residual distribution) to match the **characteristics** of your outcome variable
- Accommodates many **continuous** and **categorical** outcome variables
- Includes **logistic regression** and **Poisson regression**

3 Logistic regression

3.1 Logistic regression

3.1.1 (Binary) logistic regression

- **Outcome**: binary
  - Observed value (Y): 0 or 1, where 1 = "success" or "event"
  - Predicted value ($\hat{Y}$): **Probability** of success, between 0 and 1
- **Residual distribution**: binomial
- **Link function**: logit (or log-odds) = $ln\left(\frac{\hat{Y}}{1-\hat{Y}}\right)$

$$ln\left(\frac{\hat{Y}}{1-\hat{Y}}\right) = ln\left(\frac{\hat{p}}{1-\hat{p}}\right) = b_0 + b_1X_1 + b_2X_2 + \cdots + b_pX_p$$

3.1.2 Reminder: normal distribution

$$f(x) = \frac{1}{\sqrt{2\pi\sigma^2}}e^{-\frac{(x-\mu)^2}{2\sigma^2}}$$

Mean of normal distribution = μ

Variance of normal distribution = σ^2

- Mean and variance are **different parameters** and are **unrelated**

3.1.3 Binomial distribution

$$P(X = k) = \binom{n}{k} p^k (1-p)^{n-k}$$

- n is the sample size
- p is the probability of an event

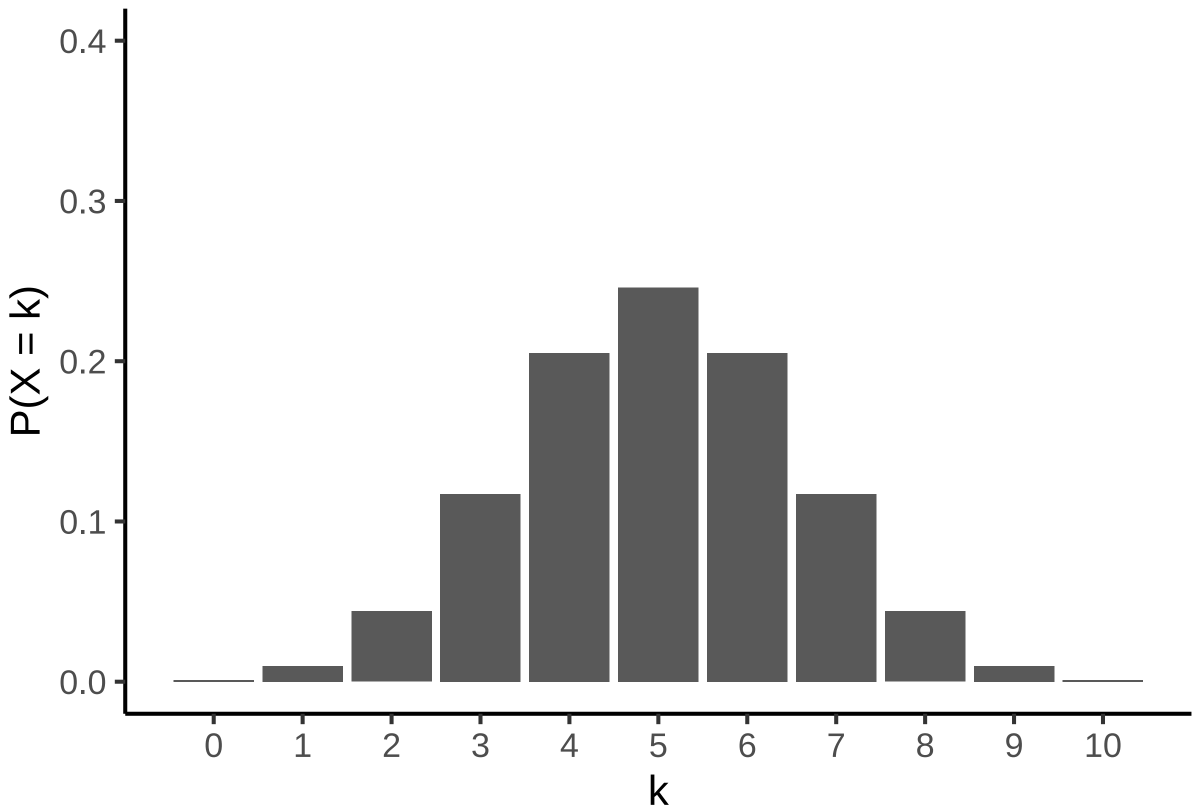
- k is the observed number of events

- $\binom{n}{k} = \frac{n!}{k!(n-k!)}$ and is read as "n choose k"

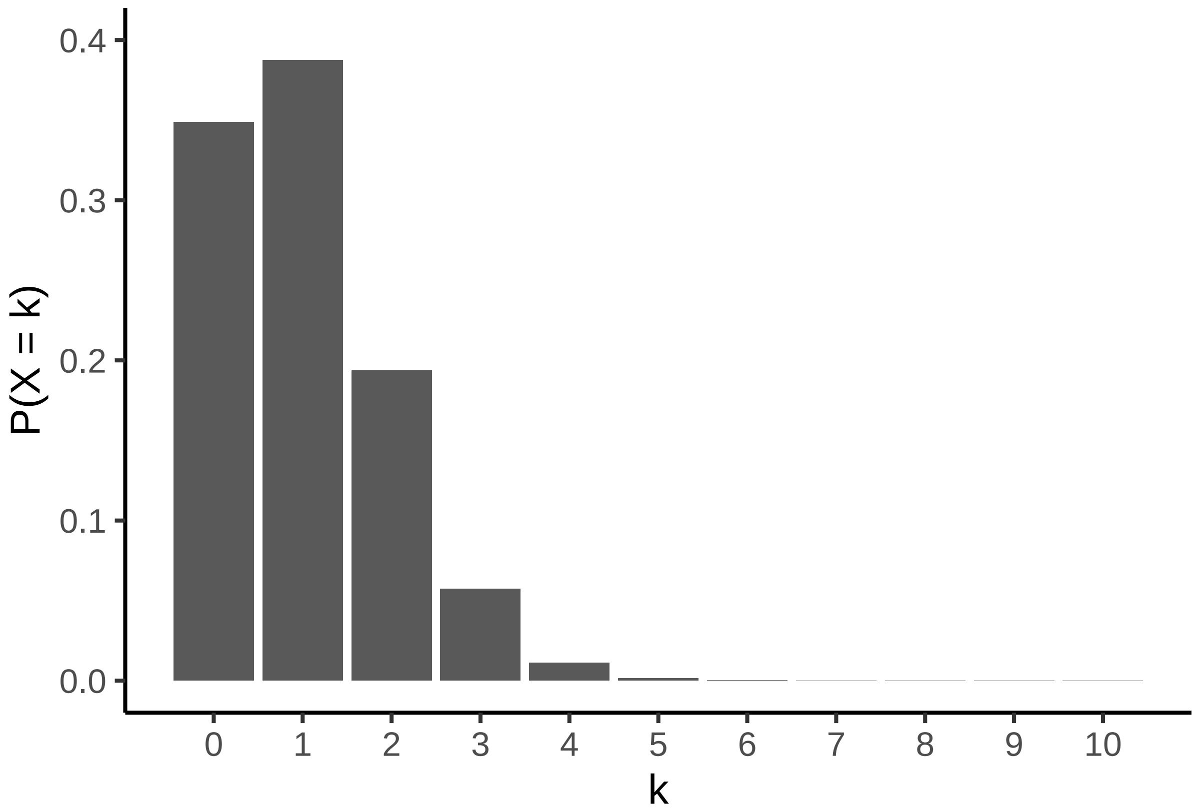
3.1.4 Binomial distribution

What is the probability of having k events in n trials, each of which has probability p of being an "event"?

- $p = 0.5$, $n = 10$



- $p = 0.1$, $n = 10$



3.1.5 Binomial distribution

$$P(X = k) = \binom{n}{k} p^k (1-p)^{n-k}$$

Mean of a binomial distribution: np
Variance of a binomial distribution: $np(1-p)$

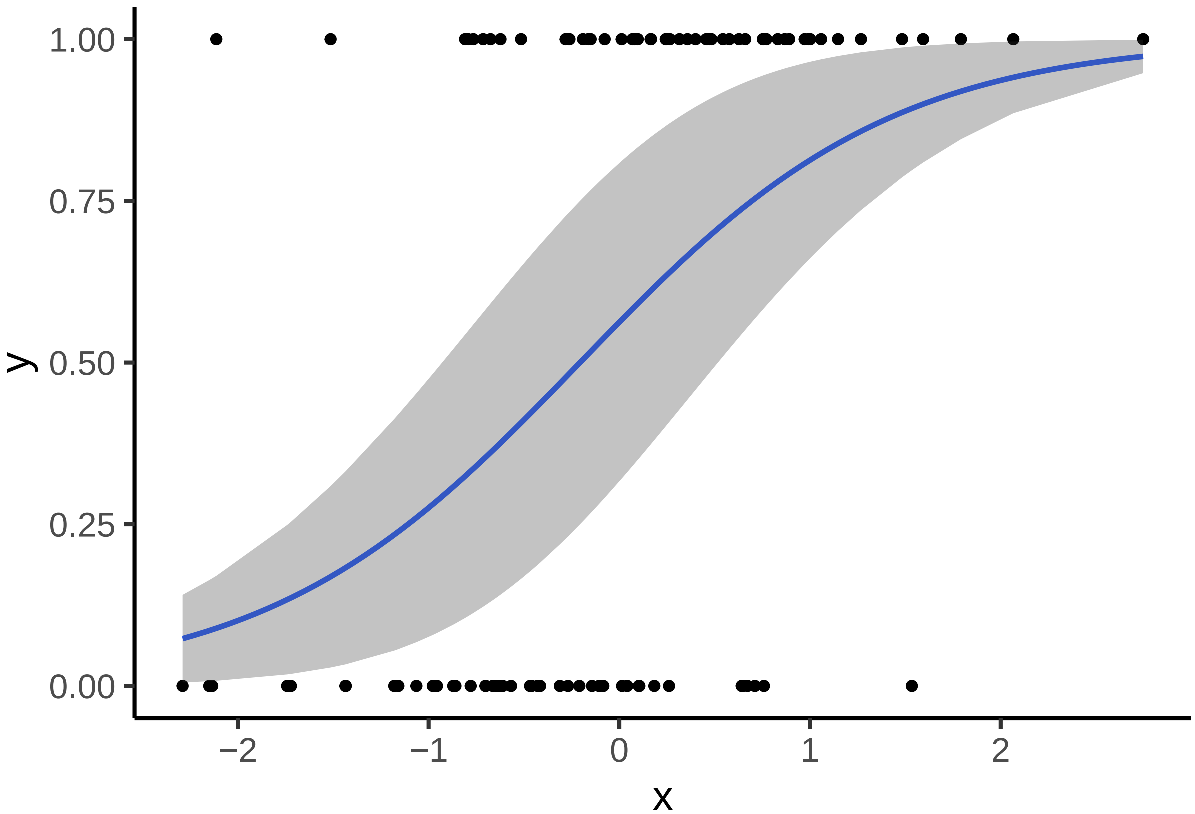
- Mean and variance are **related** to one another
  - They are functions of the same parameters (n and p)
- **Heteroscedasticity is built into logistic regression**

3.1.6 Logistic regression: What we model

- Linear regression: Model the **mean** of the outcome (conditional on predictors(s))
- Logistic regression: Model the **probability of a "success" or "event"** (conditional on predictor(s))

- From the **probability**, we can also get the **odds** of a success and the **logit or log-odds** of a success

3.1.7 Figure: What we model



3.1.8 Three forms of logistic regression

Probability:

$$\hat{p} = \frac{e^{(b_0+b_1X_1+b_2X_2+\cdots+b_pX_p)}}{1+e^{(b_0+b_1X_1+b_2X_2+\cdots+b_pX_p)}}$$

Odds:

$$\hat{odds} = \frac{\hat{p}}{1-\hat{p}} = e^{b_0+b_1X_1+b_2X_2+\cdots+b_pX_p}$$

Logit:

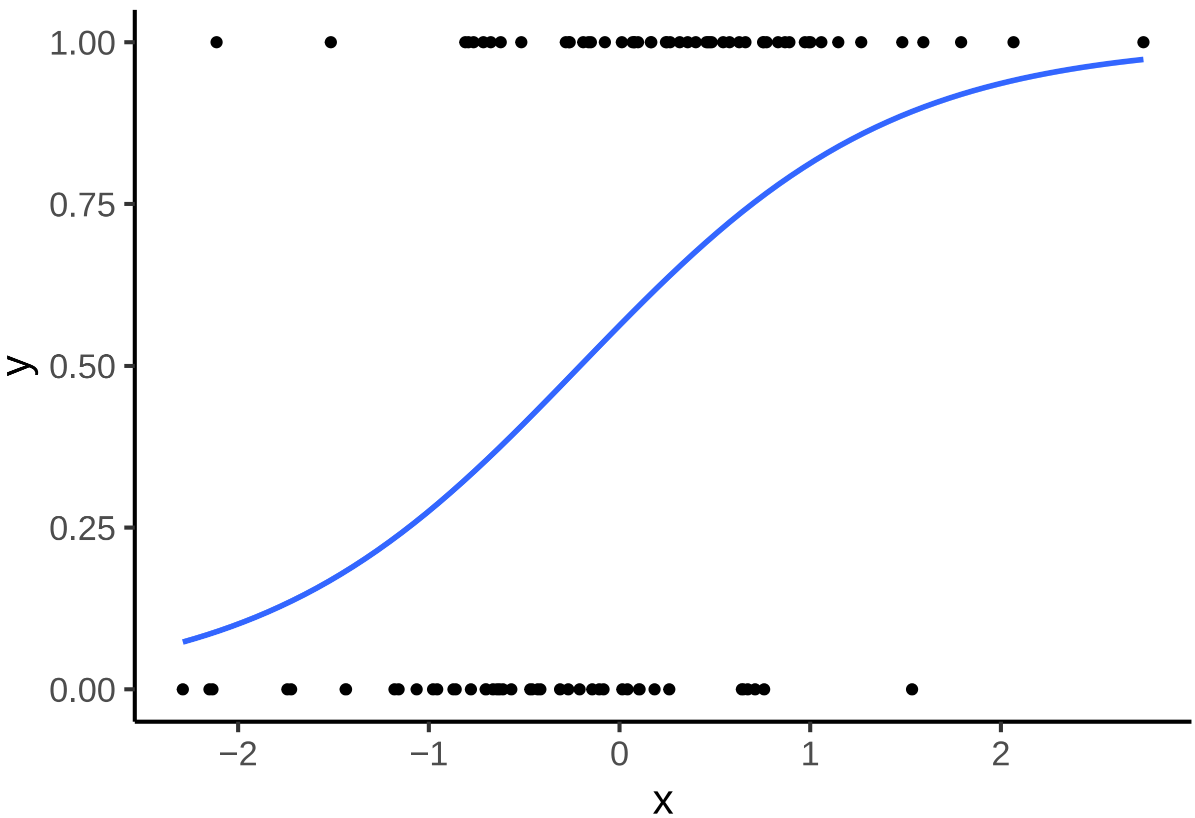
$$ln\left(\frac{\hat{p}}{1-\hat{p}}\right) = b_0 + b_1X_1 + b_2X_2 + \cdots + b_pX_p$$

3.2 Probability metric

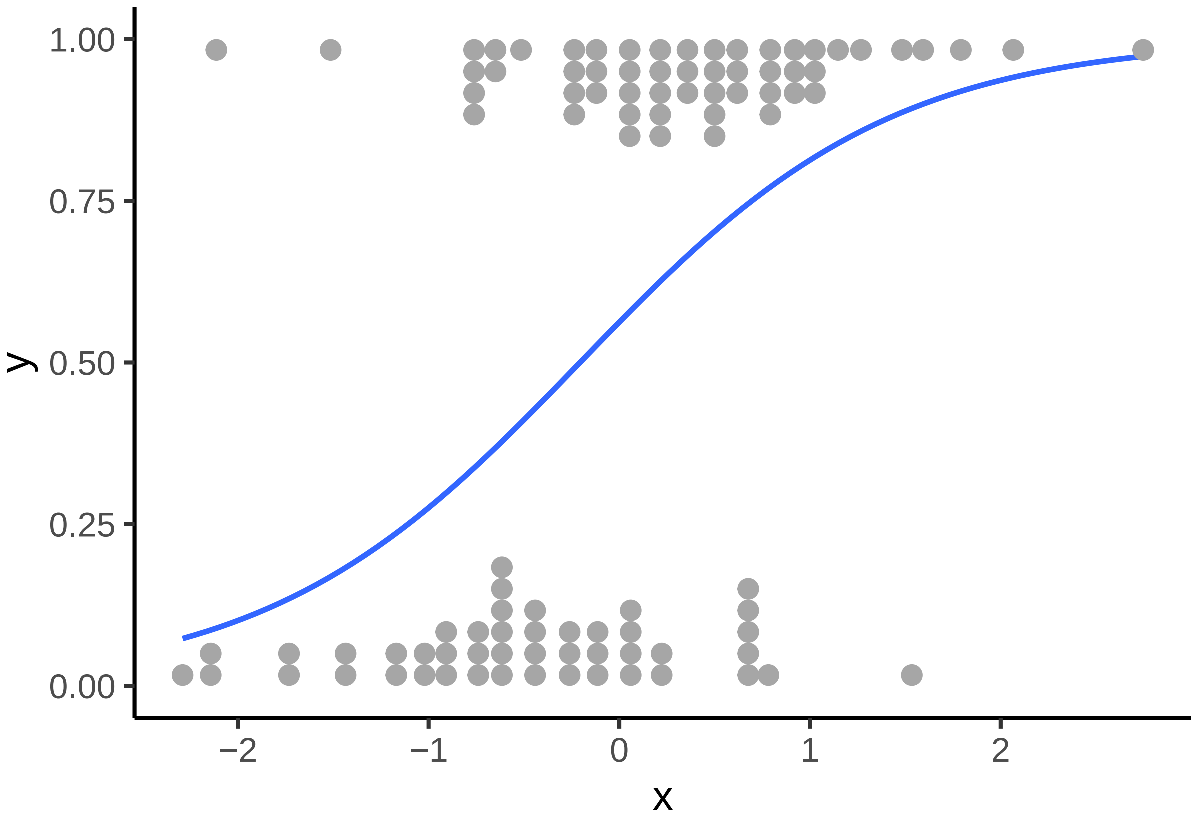
3.2.1 What is probability (p)?

- Likelihood of a "success" or "event"
- Ranges from 0 to 1
- Both options are equally likely when $p = 0.5$

3.2.2 $\hat{p} = \frac{e^{0.251+1.219X}}{1+e^{0.251+1.219X}}$



3.2.3 $\hat{p} = \frac{e^{0.251+1.219X}}{1+e^{0.251+1.219X}}$



3.2.4 Probability metric interpretation: General

$$\hat{p} = \frac{e^{0.251+1.219X}}{1+e^{0.251+1.219X}}$$

General interpretation of **intercept**:

b_0 is related to the **probability of success** when $\mathbf{X = 0}$

- $b_0 > 0$: Success (1) more likely than failure (0) when X = 0
- $b_0 < 0$: Failure (0) more likely than success (1) when X = 0

3.2.5 Probability metric interpretation: General

$$\hat{p} = \frac{e^{0.251+1.219X}}{1+e^{0.251+1.219X}}$$

General interpretation of **slope**:

b_1 tells you how **predictor X relates to probability of success**

- $b_1 > 0$: Probability of a success increases as X increases
- $b_1 < 0$: Probability of a success decreases as X increases

3.2.6 Probability metric interpretation: Example

$$\hat{p} = \frac{e^{0.251+1.219X}}{1 + e^{0.251+1.219X}}$$

Interpretation of example **intercept**:

- $b_0 > 0$: **Success (1) more likely than failure (0) when X = 0**

- Probability of success when X = 0:

$\frac{e^{b_0}}{1+e^{b_0}} = \frac{e^{0.251}}{1+e^{0.251}} = 0.562$

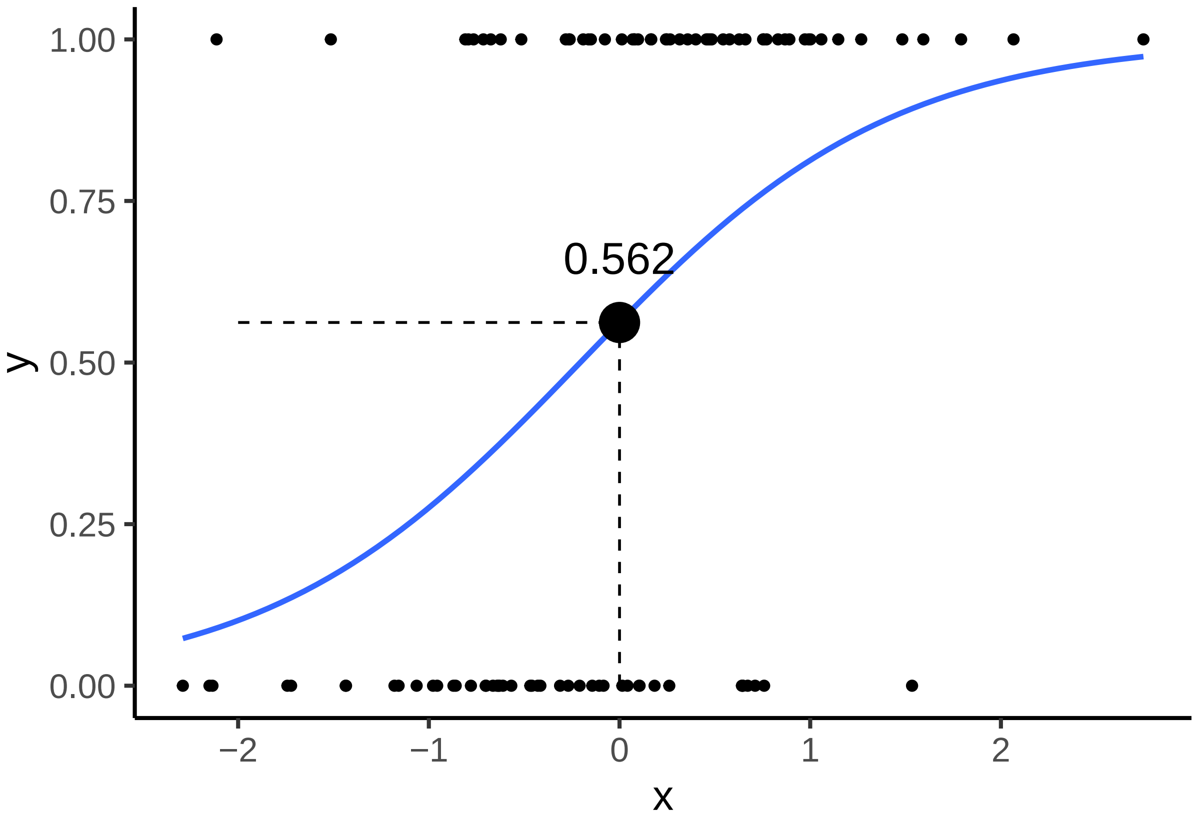
3.2.7 Probability metric interpretation: Example

$$\hat{p} = \frac{e^{0.251+1.219X}}{1 + e^{0.251+1.219X}}$$

Interpretation of example **slope**:

- $b_1 > 0$: **Probability of a success increases as X increases**

3.2.8 P(success|X=0)



3.2.9 Probability metric interpretation: Non-linear

- Linear regression:
  - **Constant**, linear slope
  - Slope depends on the **slope only**
- Logistic regression (probability):
  - **Non-linear** slope
  - Slope depends on BOTH **slope** (b_1) and **predicted probability** ($\hat{p}$)
    - The slope of the **tangent to the regression line** at the predicted outcome value $= \hat{p}(1 - \hat{p})b_1$

3.2.10 Probability metric interpretation: Non-linear

When $X = 1.5$:

$$\hat{P}(success) = \hat{p} = \frac{e^{b_0 + b_1 X}}{1 + e^{b_0 + b_1 X}} = \frac{e^{0.251 + 1.219 \times 1.5}}{1 + e^{0.251 + 1.219 \times 1.5}} = 0.889$$

Approximate **slope** at that point is

$$\hat{p}(1 - \hat{p})b_1 = 0.889 \times (1 - 0.889) \times 1.219 = 0.12$$

3.2.11 Probability metric interpretation: Non-linear

X value	Predicted probability	Slope
-3	0.03	0.04
-2	0.10	0.11
-1	0.28	0.24
0	0.56	0.30
1	0.81	0.19
2	0.94	0.07
3	0.98	0.02

3.2.12 A caution about probability equation

Warning

You might also see the probability defined as $\hat{p} = \frac{1}{1+e^{-(b_0+b_1X)}}$
Or more generally, $\hat{p} = \frac{1}{1+e^{-(Xb)}}$

- These are **numerically equivalent** to what we've talked about
  - But did you notice the negative sign?
  - No? You didn't expect it and missed it in the complicated equation?
  - Yeah, that's why we don't use this version

3.3 Odds metric

3.3.1 What are odds?

Odds is the **ratio of two probabilities**

- Model the probability of a "success"
- Odds is the ratio of probability of a "success" ($\hat{p}$) to the probability of "not a success" ($1 - \hat{p}$)

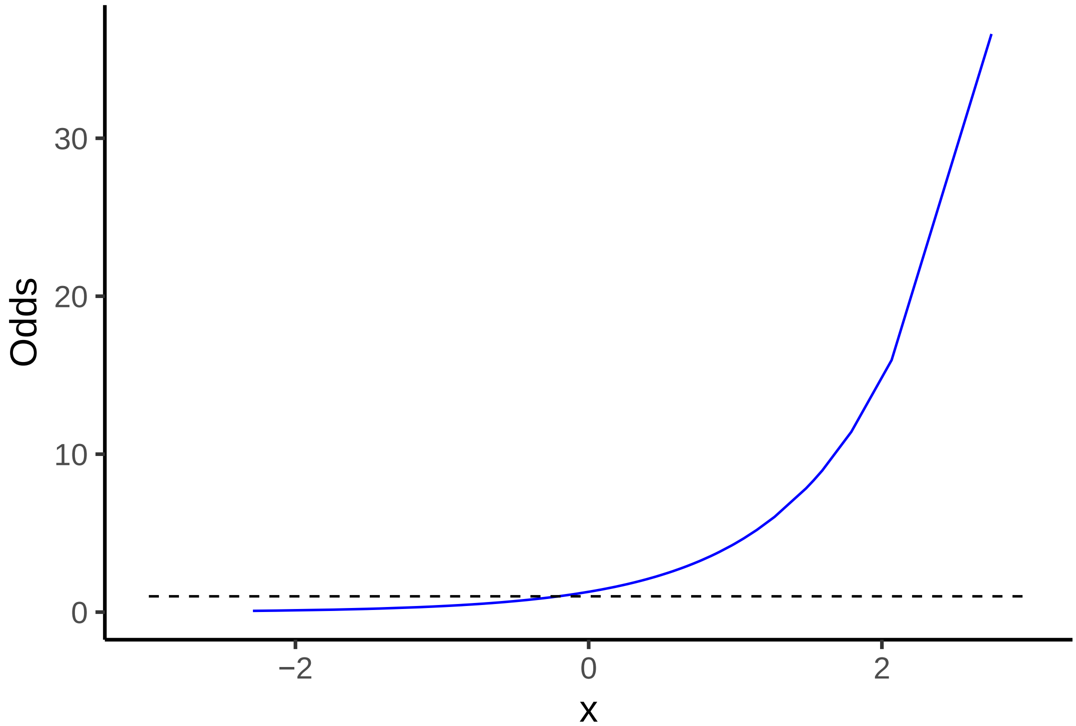
$$odds = \frac{\hat{p}}{(1 - \hat{p})}$$

As **probability** of "success" increases (nonlinearly), the **odds** of "success" increases (also nonlinearly, **but in a different way**)

3.3.2 How do odds work?

- Probability ranges from 0 to 1, switches at 0.5
  - **Success more likely** than failure when $p > 0.5$
  - **Success less likely** than failure when $p < 0.5$
- Odds range from 0 to $+\infty$, switches at 1
  - **Success more likely** than failure when $odds > 1$
  - **Success less likely** than failure when $odds < 1$

3.3.3 $\hat{odds} = \frac{\hat{p}}{(1-\hat{p})} = e^{0.251+1.219X}$



3.3.4 Odds metric interpretation: General

$$\hat{odds} = \frac{\hat{p}}{(1-\hat{p})} = e^{0.251+1.219X}$$

General interpretation of **intercept**:
b_0 is related to the **odds of success when** $X = \mathbf{0}$

- Odds of success **when X = 0**: e^{b_0}
- $b_0 > 0$: Odds of success > 1 when $X = 0$
- $b_0 < 0$: Odds of success < 1 when $X = 0$

3.3.5 Odds metric interpretation: General

$$\hat{odds} = \frac{\hat{p}}{(1-\hat{p})} = e^{0.251+1.219X}$$

General interpretation of **slope**:
b_1 = relationship between predictor X and the odds of success

- $b_1 > 0$: Odds of success **increases** as X **increases**

- $b_1 < 0$: Odds of a success **decreases** as X **increases**

3.3.6 Odds metric interpretation: Example

$$\hat{odds} = \frac{\hat{p}}{(1-\hat{p})} = e^{0.251+1.219X}$$

Interpretation of example **intercept**:

- $b_0 > 0$: Odds of success > 1 when $X = 0$
  - Success (1) more likely than failure (0) when $X = 0$

- Odds of success when $X = 0$: $e^{b_0} = e^{0.251} = 1.29$
  - A "success" is about 1.29 times as likely as a "failure"
  - Compare to 0.562 probability of success: 0.562 / 0.438 = 1.28

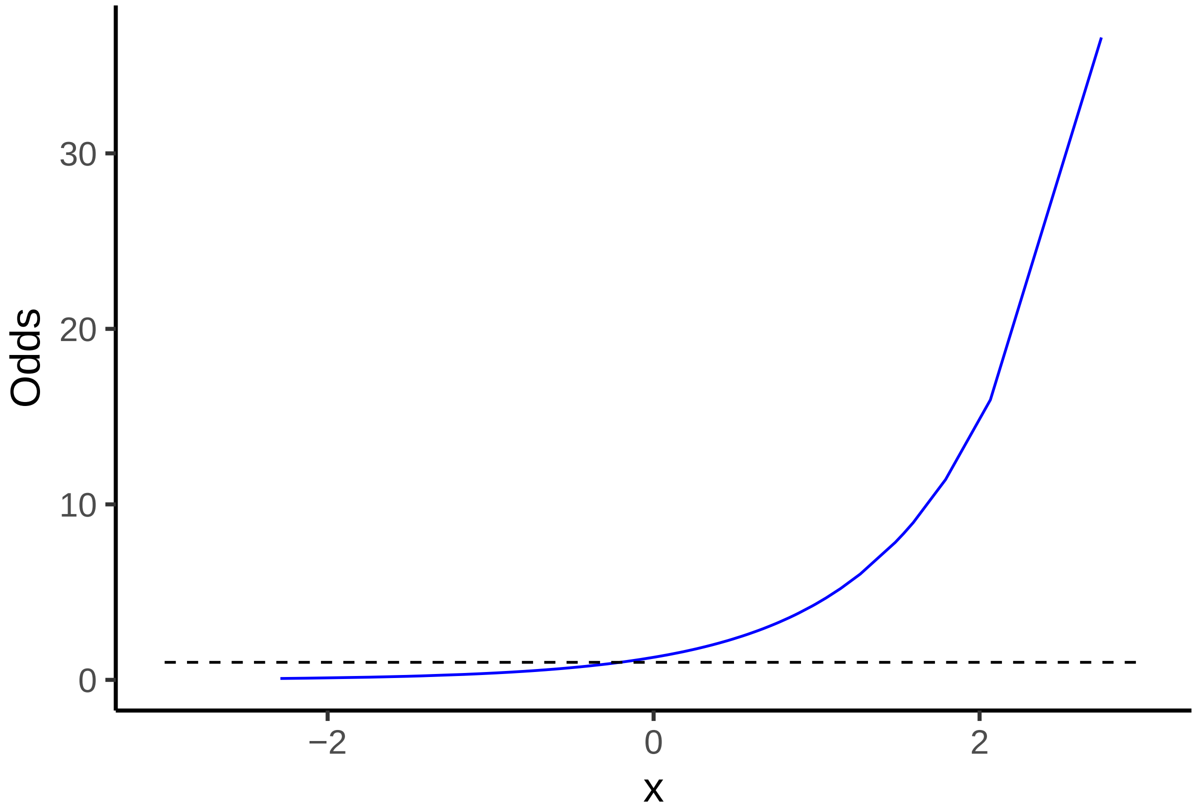
3.3.7 Odds metric interpretation: Example

$$\hat{odds} = \frac{\hat{p}}{(1-\hat{p})} = e^{0.251+1.219X}$$

Interpretation of example **slope**:

$b_1 > 0$: Odds of a success **increases** as X **increases**

3.3.8 Odds metric interpretation: Non-linear



3.3.9 Odds metric interpretation: Non-linear

- This non-linear change is presented in terms of **odds ratio**
  - Constant, **multiplicative** change in predicted odds
  - For a 1-unit difference in X, the **predicted odds** of success is **multiplied by** the **odds ratio**
- *Example*: odds ratio $= e^{b_1} = e^{1.219} = 3.38$
  - For a 1-unit difference in X, the **predicted odds** of success is **multiplied by** 3.38

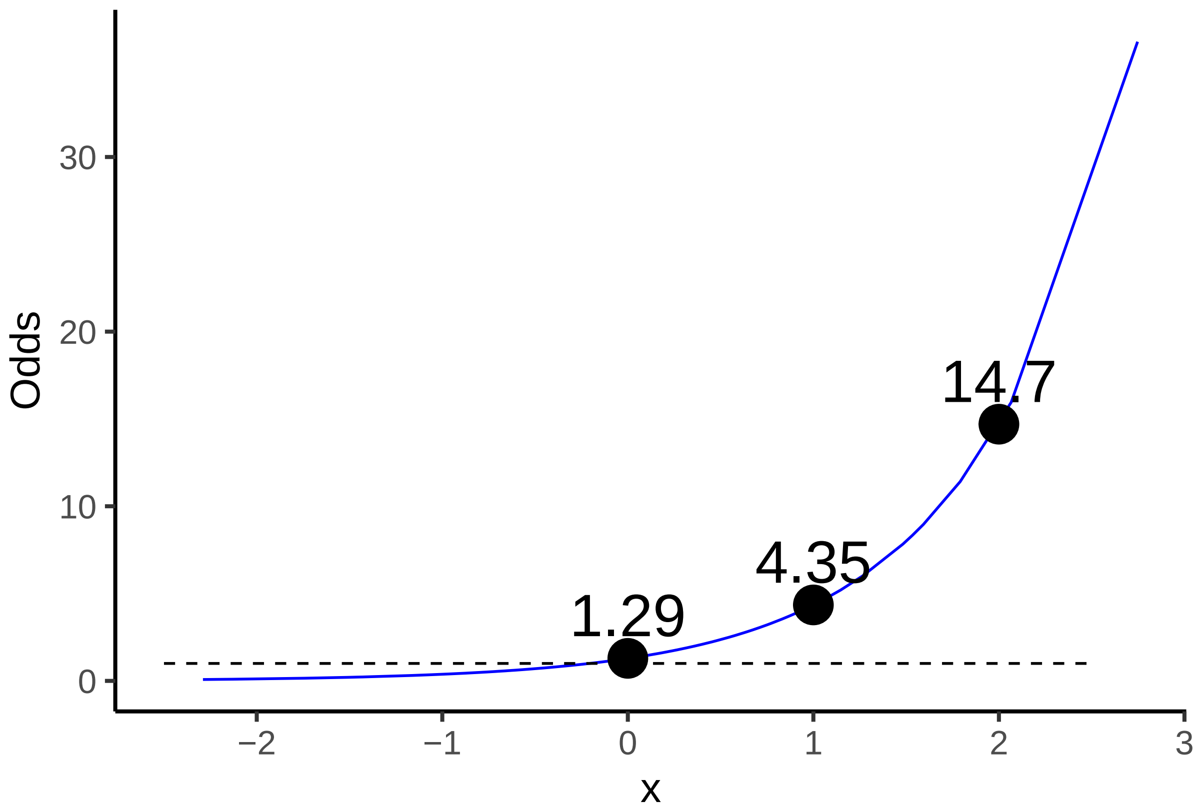
3.3.10 Odds metric interpretation: Non-linear

- Odds ratio $= e^{b_1} = e^{1.219} = 3.38$
- Odds ratio for $X = 1$ versus $X = 0$: $\frac{odds(X=1)}{odds(X=0)} = \frac{4.3492351}{1.2853101} = 3.38$

  - Odds of success is 3.38 times larger when $X = 1$ vs $X = 0$

- Odds ratio for $X = 2$ versus $X = 1$: $\frac{odds(X=2)}{odds(X=1)} = \frac{14.7169516}{4.3492351} = 3.38$
  - Odds of success is 3.38 times larger when $X = 2$ vs $X = 1$
- In fact, **ANY 1 unit difference** in X
- **Constant multiplicative** change

3.3.11 Odds metric figure again (odds ratio = 3.38)



3.3.12 Odds metric interpretation: Non-linear

X value	Predicted probability	Predicted odds
-3	0.03	0.03
-2	0.10	0.11

X value	Predicted probability	Predicted odds
-1	0.28	0.38
0	0.56	1.29
1	0.81	4.35
2	0.94	14.72
3	0.98	49.80

3.3.13 A caution about odds

Warning

- **Odds ratios** are very popular in medicine and epidemiology
- They can be **extremely** misleading
- The **same odds ratio** corresponds to many **different** probability values
  - Odds ratio $= \frac{odds=3}{odds=1} = 3$
    - Corresponds to probability of 0.75 vs 0.5
  - Odds ratio $= \frac{odds=9}{odds=3} = 3$
    - Corresponds to probability of 0.90 vs 0.75

3.4 Logit or log-odds metric

3.4.1 What is the logit?

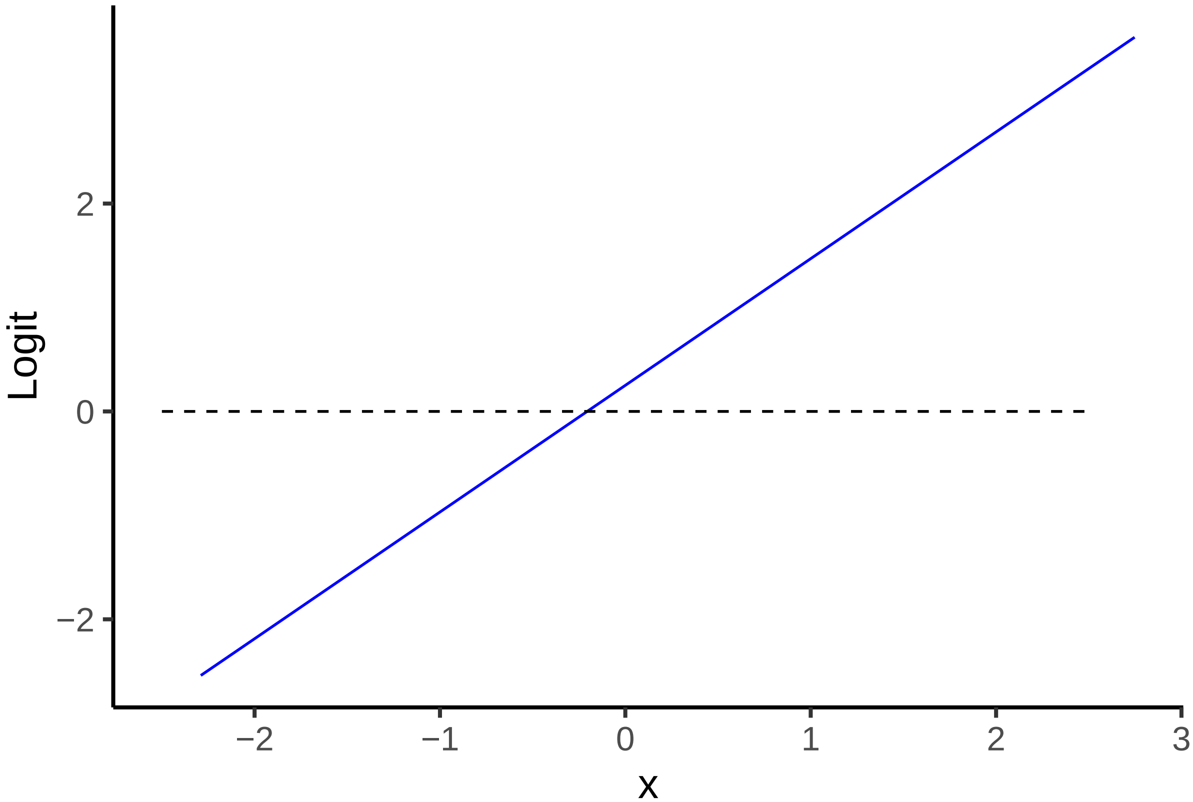
Logit or **log-odds** is the natural log (ln) of the odds

- As **probability** of "success" increases (nonlinearly, S-shaped curve)
  - The *odds* of "success" increases (also nonlinearly, exponentially up)
  - The **logit** of "success" increases **linearly**

3.4.2 How does the logit work?

- Probability ranges from 0 to 1, **switches at 0.5**
- Odds range from 0 to $+\infty$, **switches at 1**
- Logit ranges from $-\infty$ to $+\infty$, **switches at 0**
  - Success more likely than failure when logit > 0
  - Success less likely than failure when logit < 0

3.4.3 $\hat{logit} = ln\left(\frac{\hat{p}}{(1-\hat{p})}\right) = 0.251 + 1.219X$



3.4.4 Logit metric interpretation: General

$$\hat{logit} = ln\left(\frac{\hat{p}}{(1-\hat{p})}\right) = 0.251 + 1.219X$$

General interpretation of **intercept**:
b_0 is related to the logit of success when X = 0

- Logit of success when X = 0: b_0
- $b_0 > 0$: Logit > 0 when X = 0
- $b_0 < 0$: Logit < 0 when X = 0

3.4.5 Logit metric interpretation: General

$$\hat{logit} = ln\left(\frac{\hat{p}}{(1-\hat{p})}\right) = 0.251 + 1.219X$$

General interpretation of **slope**:
b_1 is the relationship between predictor X and logit of success

- $b_1 > 0$: Logit of a success increases as X increases
- $b_1 < 0$: Logit of a success decreases as X increases

3.4.6 Logit metric interpretation: Example

$$\hat{logit} = ln\left(\frac{\hat{p}}{(1-\hat{p})}\right) = 0.251 + 1.219X$$

Interpretation of example **intercept**

- $b_0 > 0$: Logit > 0 when X $= 0$
- Logit of success when X $= 0$: $b_0 = 0.251$

3.4.7 Logit metric interpretation: Example

$$\hat{logit} = ln\left(\frac{\hat{p}}{(1-\hat{p})}\right) = 0.251 + 1.219X$$

Interpretation of example **slope**

- $b_1 > 0$: Logit of a success increases by 1.219 units when X increases by 1 unit

3.5 Metrics wrap-up

3.5.1 So which metric should I use?

They are **equivalent**, so use the metric that

- Makes the most *sense* to you
- You can *explain* fully
- Is most commonly used in your *field*

3.5.2 Some things to keep in mind

- Odds ratios tell you about **change**, but not **where you start**
  - If you report odds ratios, *also report some measure of probability* e.g., probability of success at the mean of X
  - 10x change is 5 to 50 or 0.05 to 0.5?
- Logit is nice because it's **linear**, but it's not very **interpretable**
  - What is a "logit"? It's just a *mathematical* concept that makes a straight line – not actually **meaningful**
  - *But many psychology measures don't have meaningful metrics...*

3.5.3 Confidence intervals

Default results are in **logit metric**: compare to **null value** of 0

term	estimate
(Intercept)	0.251
x	1.219

Confidence intervals are in **logit metric**: does it contain 0?

	2.5 %	97.5 %
(Intercept)	-0.188	0.703
x	0.661	1.876

3.5.4 Confidence intervals

$e^{estimate}$ converts to **odds ratio metric**: compare to **null value** of 1

term	estimate	OR
(Intercept)	0.251	1.285
x	1.219	3.383

$e^{estimate}$ converts to **odds ratio metric**: does it contain 1?

	2.5 %	97.5 %	OR 2.5 %	OR 97.5 %
(Intercept)	-0.188	0.703	0.829	2.019
x	0.661	1.876	1.938	6.528

3.6 A tiny detour

3.6.1 Three alternatives / extensions

- What if I want to focus more on **probability** (and don't care about odds ratios)?
  - Probit regression: based on the cumulative normal distribution, not the logistic distribution
- What if I have **three or more** options for my outcome?
  - Categories have an order to them: Ordinal logistic regression
  - Categories have no order to them: Multinomial logistic regression

4 Estimation and model fit

4.1 Estimation

4.1.1 You ran a model: What now?

Usually two things you want to do with it

- Compute some measure of predictive power or **model fit**
  - $R^2_{multiple}$ or similar
- **Compare** that model to another competing model
  - Which model is better?

4.1.2 Model estimation

Linear regression is estimated using **ordinary least squares** (OLS)

- Produces sums of squares (SS)
- Measures like R^2 are a function of SS

GLiMs (like logistic regression) are estimated using **maximum likelihood**

- No sums of squares
- Instead: **Deviance**, which is a function of the *log-likelihood*

4.1.3 What is deviance?

- Conceptually similar to $SS_{residual}$
- If you had n predictors
  - One predictor per person
  - Perfectly predict the outcome values
  - "Perfect" model
- **Deviance** is how far from this "perfect" model you are
  - This is "badness" of fit

4.2 R^2 measures

4.2.1 R^2 in linear regression

- R^2 for linear regression has many **desirable qualities**
  - Always ranges from 0 to 1
  - Always stays the same or increases with more predictors (never decreases)

Without $SS_{residual}$, what can we do?

4.2.2 R^2 analogues

- There are some **general measures** that work for all GLiMs and some more **specific measures** that only work for *logistic regression*

> ⚠ Warning
>
> R^2 analogues don't have the properties that R^2 in linear regression does
>
> - Can be less than 0 or greater than 1
> - Can decrease when you add predictors

4.2.3 Pseudo-R^2 or $R^2_{deviance}$

$$R^2_{deviance} = 1 - \frac{deviance_{model}}{deviance_{intercept.only.model}}$$

- Compare your model to a model with no predictor (only intercept)
  - Common for many types of *advanced modeling*, could do it for linear regression but probably never would
  - Essentially tests how much closer the model is to the "perfect" model than the intercept only model
  - Theoretically bounded by 0 and 1, but in practice…

4.2.4 $R^2_{McFadden}$

$$R^2_{McFadden} = 1 - \frac{LL_{model}}{LL_{intercept.only.model}}$$

- Same idea as $R^2_{deviance}$, just using LL instead of deviance
  - Theoretically bounded by 0 and 1
  - Relatively independent of **base rate**
    - **Base rate** is the overall **probability of a success** in the sample
    - See DeMaris (2002) for more details about logistic regression specific measures

4.2.5 R^2 as correlation between observed and predicted values

- In linear regression, $R^2_{multiple}$ is *also* the squared correlation between the **observed** Y **values** and the **predicted** Y **values**
- Most software packages can produce *predicted* Y *values* for your analysis
  - Save predicted values to the dataset
  - Correlate **observed** and **predicted** Y values (squared correlation)

4.3 Model comparisons

4.3.1 Model comparisons

- In linear regression, if you **added a predictor**, there were two ways to tell if that predictor was adding to the model:
  - Test of the **regression coefficient** (i.e., Wald test: t-test or z-test)
  - R^2_{change} for added prediction (with its F-test)
- **For logistic regression, Wald test of the regression coefficient may not be reliable** (see Vaeth, 1985)
  - Need to use some analogue of the *significance test* for R^2_{change}

4.3.2 Likelihood ratio (LR) test

- **Ratio** of **likelihoods**
  - Specifically, a **function of likelihood** from ML estimation
  - Even more specifically, $-2 \times log - likelihood$
  - $-2 \times LL$ is the **deviance**
- Test statistic
  - $\chi^2 = deviance_{model1} - deviance_{model2}$
  - How did we get from **ratio** to **difference**?
    - *Division* in *log* metric is *subtraction* in *regular* metric

4.3.3 Likelihood ratio (LR) test

$$\chi^2 = deviance_{model1} - deviance_{model2}$$

- Model 1: simpler model (fewer predictors, worse fit)
- Model 2: more complex model (more predictors, better fit)
- **Degrees of freedom** = difference in number of parameters
  - **Significant test**: Model 1 is significantly worse than Model 2
  - **NS test**: Model 1 and 2 are not significantly different, so go with simpler one (Model 1)

4.3.4 LR test: Example

- Logistic regression example: Deviance = 116.146
- Logistic regression model with no predictors (intercept only): Deviance = 137.989
- $\chi^2(1) = 137.989 - 116.146 = 21.843$
  - Critical value for χ^2 with 1 df and $\alpha = 0.05$ is 3.841
  - The test is significant: $21.843 > 3.841$
    - Model 2 is better than Model 1
    - The predictor is significant

5 Summary

5.1 Summary

5.1.1 Summary

- Use logistic regression when your outcome is binary
  - **Don't use linear regression**
- Be careful with interpretation no matter what
  - **Probability**: Probability *makes sense*, but it's *nonlinear*
  - **Odds**: Odds ratio *seems to make sense* but it can be *misleading*
  - **Logit**: *Linear* but what even is a *logit*?
- But many basic concepts parallel linear regression

  - Intercept, slope(s), linear combination, $R^2_{multiple}$

5.1.2 In class

- We will
  - Run some logistic regression models
  - Interpret the results