

# Lecture 9: Expanding binary logistic regression

PSYC234: Statistics: from association to modelling causality

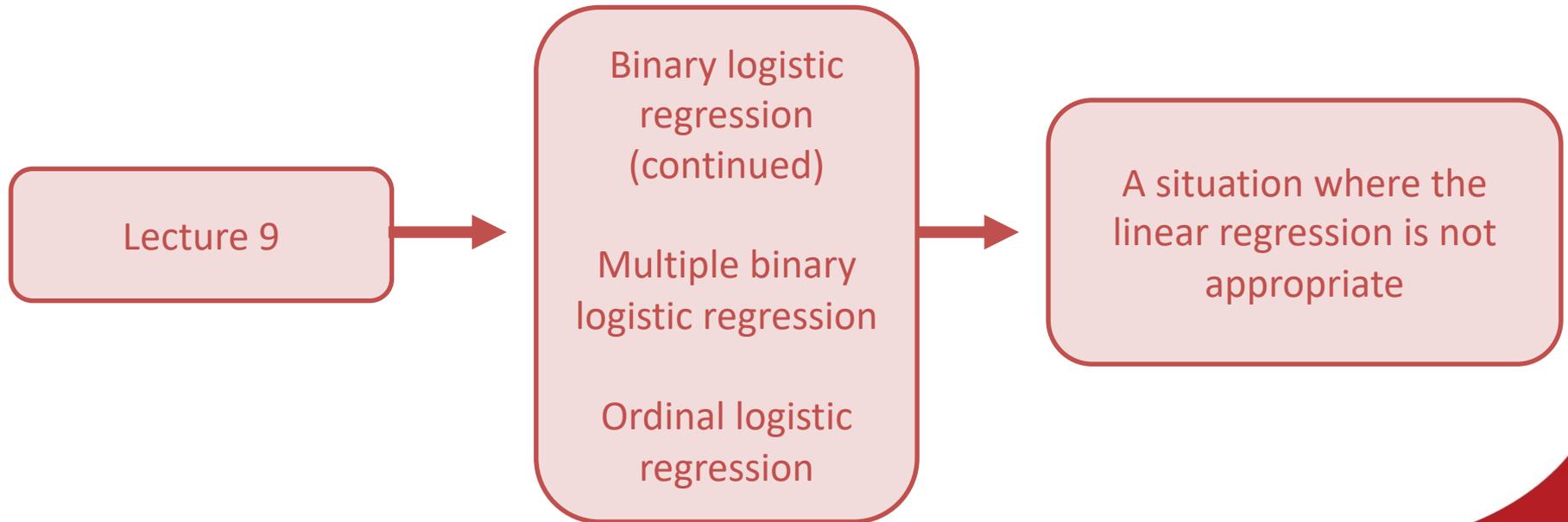
Dr Amy Atkinson

Lecturer in Developmental Psychology

[amy.atkinson@lancaster.ac.uk](mailto:amy.atkinson@lancaster.ac.uk)

# The plan

**My aim:** to add a few final statistical tests to your toolbox for when the statistical test you've learned about might not be appropriate



# Learning objectives

---

- To understand how to conduct binary logistic regression models in R with categorical predictors (3+ levels) and continuous predictors, and interpret the output
- To understand how to conduct multiple logistic regression models in R and interpret the output
- To understand how to conduct ordinal logistic regression models in R and interpret the output

## Part 1

Binary logistic regression with other types of predictors

# Categorical predictors with more than two levels

# Interpretation

---

- Set one level as your reference category that all other levels are compared to (e.g. “Single”)
- Everything else is the same

# 5. Evaluating individual predictors

## The intercept:

Log odds of being happy (happy = yes) in the reference category (i.e. Single)

Coefficients:

	Estimate	Std. Error	z value	Pr(> z )
(Intercept)	-0.4055	0.5270	-0.769	0.4417
marital_statusCohabiting	0.9445	0.7099	1.330	0.1834
marital_statusMarried	1.8524	0.7659	2.419	0.0156 *
marital_statusDivorced	-1.0415	0.7659	-1.360	0.1739

---

Signif. codes: 0 '\*\*\*' 0.001 '\*\*' 0.01 '\*' 0.05 '.' 0.1 ' ' 1

## 5. Evaluating individual predictors

### Single vs cohabiting:

The change in log odds of being happy (happy = yes) when going from Single to Cohabiting

Coefficients:

	Estimate	Std. Error	z value	Pr(> z )
(Intercept)	-0.4055	0.5270	-0.769	0.4417
marital_statusCohabiting	0.9445	0.7099	1.330	0.1834
marital_statusMarried	1.8524	0.7659	2.419	0.0156 *
marital_statusDivorced	-1.0415	0.7659	-1.360	0.1739

---

Signif. codes: 0 '\*\*\*' 0.001 '\*\*' 0.01 '\*' 0.05 '.' 0.1 ' ' 1

# Continuous predictors

# What's different for continuous predictors?

---

The process is the same except for:

1. Predictor should be a numeric or integer variable (instead of a factor)
2. Way to check quasi-complete separation and complete separation
3. Interpretation of the Estimates and odds ratios
4. An additional assumption: linearity of the logit

# Example

## Hours free time and happiness

Does the number of hours free time an individual has predict response to the following survey question:

Are you happy?

- Yes
- No



- Predictor: Hours free time (continuous)
- Outcome: Happiness (Yes/No)

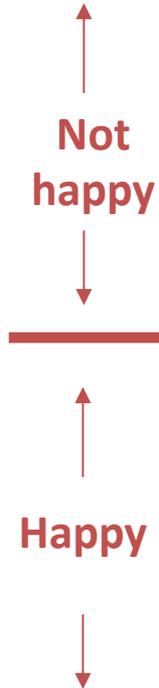
# 1. Is the predictor variable a numeric/integer?

---

```
> str(hours_data$Hours_free_time)
int [1:53] 19 19 15 18 15 17 18 13 12 10 ...
```

## 2. Complete separation and quasi-complete separation

Participant_ID	Hours_free_time	Happy	Happy_numeric
1	10	No	0
2	17	No	0
3	18	No	0
4	9	No	0
5	11	No	0
6	8	No	0
7	19	No	0
8	3	No	0
9	5	No	0
10	14	No	0
11	6	Yes	1
12	12	Yes	1
13	15	Yes	1
14	20	Yes	1
15	4	Yes	1
16	7	Yes	1
17	1	Yes	1
18	2	Yes	1
19	13	Yes	1
20	16	Yes	1



Complete separation: The predictor perfectly predicts the outcome

Hours\_free\_time perfectly predicts Happy

When hours\_free\_time is 15 or below, Happy = No

When hours\_free\_time is above 15, Happy = Yes

# Warning messages

Model did not  
converge – ignore  
output!!!

You generally get warning messages when  
there is complete separation and the culprit  
predictor is continuous

```
> model1 <- glm(Happy_numeric ~ Hours_free_time, data = data, family=binomial())
```

Warning messages:

- 1: glm.fit: algorithm did not converge
- 2: glm.fit: fitted probabilities numerically 0 or 1 occurred

Explicitly tells you you  
have complete  
separation

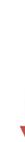
# Quasi-complete separation

Participant_ID	Hours_free_time	Happy	Happy_numeric
1	10	No	0
2	17	No	0
3	18	No	0
4	9	No	0
5	11	No	0
6	8	No	0
7	19	No	0
8	3	No	0
9	5	No	0
10	14	Yes	1
11	6	Yes	1
12	12	Yes	1
13	15	Yes	1
14	20	Yes	1
15	4	Yes	1
16	7	Yes	1
17	1	Yes	1
18	2	Yes	1
19	13	Yes	1
20	16	Yes	1

↑  
**Not happy**



↑  
**Happy**



Quasi-complete separation:  
The predictor nearly perfectly predicts the outcome

Hours\_free\_time nearly perfectly predicts Happy:

- When hours\_free\_time is below 15, Happy = No
- When hours\_free\_time is above 15, Happy = Yes
- When hours\_free\_time = 15, Happy = Yes or No

# Quasi-complete separation

## Error messages

---

```
> model2 <- glm(Happy_numeric ~ Hours_free_time, data = data2, family=binomial())  
Warning message:  
glm.fit: fitted probabilities numerically 0 or 1 occurred  
/
```

We don't get the warning message about convergence this time

But, we still get this warning message about separation

## 3. Interpretation

### We've run a model

- We've prepared our dataset, checked for separation issues, and run our model:

```
Call:
glm(formula = Happy_numeric ~ Hours_free_time, family = binomial(),
     data = hours_data)
```

```
Deviance Residuals:
    Min       1Q   Median       3Q      Max
-2.4816 -0.1619  0.1325  0.3069  1.3358
```

```
Coefficients:
            Estimate Std. Error z value Pr(>|z|)
(Intercept)  -6.0267    1.9670  -3.064 0.002185 **
Hours_free_time  0.5662    0.1635   3.464 0.000533 ***
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

(Dispersion parameter for binomial family taken to be 1)

```
Null deviance: 70.252 on 52 degrees of freedom
Residual deviance: 23.852 on 51 degrees of freedom
AIC: 27.852
```

```
Number of Fisher Scoring iterations: 6
```

The log odds of  
happy = yes when  
Hours\_free\_time  
equals 0

The change in log  
odds of happy =  
yes after a one  
unit change in  
hours\_free\_time  
(e.g. going from  
0-1 hour, or 4-5  
hours)

### 3. Interpretation Odds ratios

```
hours_model_exponentiated <- exp(hours_model$coefficients)
hours_model_exponentiated
```

The odds of  
happy = yes when  
Hours\_free\_time  
equals 0

```
(Intercept) Hours_free_time
0.002413326    1.761529986
```

The change in odds  
of happy = yes  
after a one unit  
change in the  
predictor (e.g. 0-1,  
4-5 hours)

# Assumptions

## 1. Independence of errors

- Cases of data should not be related
- For instance, each cases should represent data from a different person

We can't really test for this – we should just know this is true based on the methodology



# Assumptions

## 2. Linearity of the logit

---

- There is a linear relationship between any continuous predictor and the log of the outcome variable

Assumption for  
continuous  
predictors only

We can test this assumption after we've run our model

## Assessing assumption 2: The linearity of the logit

- This can be tested by looking at whether there is a significant interaction between the predictor and its log transformation

```
hours_data$log_Hours_free_time_int <- log(hours_data$Hours_free_time)*hours_data$Hours_free_time
```

Adds a new  
column to  
hours\_data

Log of  
Hours\_free\_time  
multiplied by  
Hours\_free\_time

## Assessing assumption 2: The linearity of the logit

---

- Run a model including the original predictor and the new log\_Hours\_free\_time\_int predictor

```
hours_model2 <- glm(Happy_numeric ~ Hours_free_time + log_Hours_free_time_int, data = hours_data, family=binomial())  
summary(hours_model2)
```

# Assessing assumption 2: The linearity of the logit

```
> summary(hours_model2)
```

```
Call:
glm(formula = Happy_numeric ~ Hours_free_time + log_Hours_free_time_int,
     family = binomial(), data = hours_data)
```

Deviance Residuals:

Min	1Q	Median	3Q	Max
-2.41283	-0.04112	0.21167	0.33456	1.25956

Coefficients:

	Estimate	Std. Error	z value	Pr(> z )
(Intercept)	-13.3547	11.3350	-1.178	0.239
Hours_free_time	2.8021	3.1063	0.902	0.367
log_Hours_free_time_int	-0.6452	0.8656	-0.745	0.456

(Dispersion parameter for binomial family taken to be 1)

Null deviance: 70.252 on 52 degrees of freedom  
 Residual deviance: 23.127 on 50 degrees of freedom  
 AIC: 29.127

Number of Fisher Scoring iterations: 8

- **Do not interpret this output!!!**
- All we are interested in is whether the “log\_Hours\_free\_time\_int” variable is significant:
  - If  $p \leq 0.05$ , the variable violates this assumption
  - If  $p > 0.05$ , the variable does not violate this assumption

log\_Hours\_free\_time\_int not significant – assumption not violated