SOC542 Statistical Methods in Sociology II

Binary outcomes II

Thomas Davidson

Rutgers University

March 27, 2023

Course updates

- Homework 3 is due 3/31 at 5pm
- Projects: Data cleaning and descriptive analyses, preliminary regression models

Plan

- Interaction terms and logistic regression
- Predictions
- Marginal effects

Logistic regression refresher

Binary outcomes and logistic regression

- We are continuing to consider binary outcome variables, focusing mostly on logistic regression:

$$p_i = logit^{-1}(\beta_0 + \beta_1 x_{1i} + \beta_2 x_{1i} + ... + \beta_k x_{ki})$$

$$= \frac{1}{1 + e^{-(\beta_0 + \beta_1 x_{1i} + \beta_2 x_{1i} + ... + \beta_k x_{ki})}}$$

- The goal is to estimate p_i, the probability that the outcome $y = 1$ as a function of covariates.
- Logistic regression is a generalized linear model, where a link function is used to project a linear model onto a non-linear outcome.

Logistic regression refresher

Binary outcomes and logistic regression

- The β coefficients in a logistic regression are *log-odds*.
- $exp(\beta)$ can allows us to interpret these coefficients as *odds-ratios*.
- $\beta_x/4$ provides an upper-bound for the effect of a unit-change in x on p_i.
- We can use models to obtain *predicted probabilities*.

Interaction terms

Specifying an interaction

- If we expect there to be an **interaction** between x and z, such that the effect of x on y varies according to the level of z, we can add an **interaction term** into our model formula.

$$y = \beta_0 + \beta_1 x + \beta_2 z + \beta_3 xz + u$$

- β_1 and β_2 are now considered as the **main effects**.
- β_3 is the coefficient for the interaction term, representing the effect of x times z.

Interaction terms

Specifying an interaction

- If we're estimating an LPM we can use the standard formula as above.
- For a logistic regression, we specify an interaction in the same way within the link function:

$$P(y = 1) = p = logit^{-1}(\beta_0 + \beta_1 x + \beta_2 z + \beta_3 xz)$$

Data

Diffusion of Microfinance[1]

- Survey data from 75 villages in Karnataka, India
  - Focus only on women aged 18-65 and 72 villages
  - Listwise deletion used to drop respondents missing key variables
  - N = 8976
- Dependent variable:
  - Membership in a micro-finance Self-Help Group (SHG), N = 3357
- Independent variables:
  - Age (continuous)
  - Nativity (dummy), 72% of women not born in current village due to marriage-related migration

[1] Data from Banerjee, A., A. G. Chandrasekhar, E. Duflo, and M. O. Jackson. 2013. "The Diffusion of Microfinance." *Science* 341 (6144): 1236498–1236498. Link to paper. Harvard Dataverse link

Interaction terms

Data exploration

There are two different factors that will be useful for understanding the results. First, nonnative respondents (typically married women due to village exogamy) and SHG participants tend to be older than natives and non-participants.

```
data %>% group_by(nonnative) %>% summarize(mean(age), median(age))
```

```
## # A tibble: 2 x 3
##   nonnative `mean(age)` `median(age)`
##       <dbl>       <dbl>         <dbl>
## 1         0        30.5            28
## 2         1        36.7            35
```

Interaction terms

Data exploration

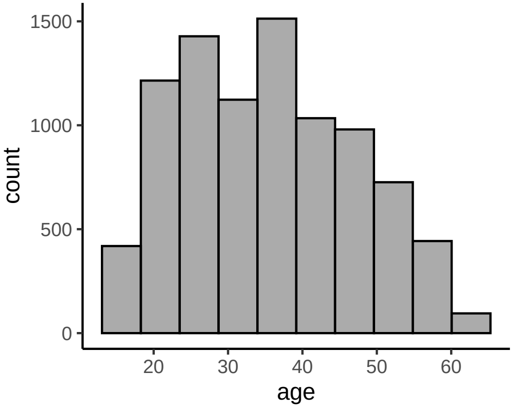
There are two different factors that will be useful for understanding the results. First, nonnative respondents (typically married women due to village exogamy) and SHG participants tend to be older than natives and non-participants.

```
data %>% group_by(shg) %>% summarize(mean(age), median(age))
```

```
## # A tibble: 2 x 3
##     shg `mean(age)` `median(age)`
##   <dbl>       <dbl>         <dbl>
## 1     0        34.1            32
## 2     1        36.4            35
```

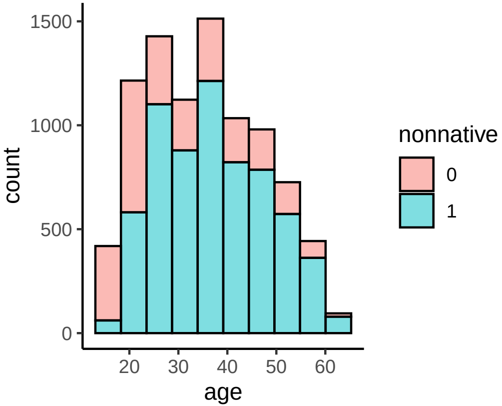
Interaction terms

Data exploration



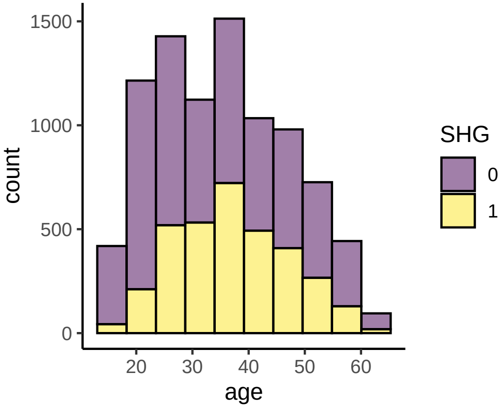
Interaction terms

Data exploration



Interaction terms

Data exploration



Interaction terms

Data exploration

Second, ~40% of nonnative women participate in SHGs, compared to only ~30% of natives.

```
data %>% group_by(nonnative, shg) %>%
    summarize(count = n(), .groups = "keep") %>% kable()
```

nonnative	shg	count
0	0	1768
0	1	751
1	0	3865
1	1	2592

Interaction terms

Estimating models

A LPM and logistic regression are used to estimate the probability of SHG membership as a function of age and nativity (whether a respondent was born in their current village of residence).

```
lpm <- lm(shg ~ age + nonnative + age:nonnative,
          data = data)
logistic <- glm(shg ~ age + nonnative + age:nonnative,
                data = data, family = binomial())
```

Interaction terms

Comparing models

	LPM	Logistic	Odds-ratio
age	0.009***	0.042***	1.043***
	(0.001)	(0.004)	(0.004)
nonnative	0.337***	1.615***	5.030***
	(0.034)	(0.159)	(0.800)
age $\times$ nonnative	-0.008***	-0.037***	0.963***
	(0.001)	(0.004)	(0.004)
Num.Obs.	8976	8976	8976
Log.Lik.	-6109.447	-5818.085	-5818.085
F	69.564	64.991	64.991

Note: ^^ * $p < 0.05$, ** $p < 0.01$, *** $p < 0.001$

Interaction terms

Intepretations

- In both models, the coefficients for the main effects of age and nativity are positive.
- The coefficients for interaction terms are both negative.
  - This implies that there is a negative effect of age for nonnative women. In other words, as age increases the probability of belonging to an SHG decreases.
- However, it is difficult to understand these interactions by only considering the coefficients, since the relationship between variables in a logistic regression is non-linear.

Predictions

Understanding interactions using predictions

- One of the ways we can start to make sense of these interactions is by making predictions.
- Let's consider predictions for a nonnative woman aged 25:

```
c1 <- coefficients(lpm)
c2 <- coefficients(logistic)
p.lpm <- as.numeric(c1[1] + c1[2]*25 + c1[3] + c1[4]*25)
print(p.lpm)
```

```
## [1] 0.3885258
```

```
p.logit <- invlogit(as.numeric(c2[1] + c2[2]*25 + c2[3] + c2[4]*25))
print(p.logit)
```

```
## [1] 0.3885585
```

Predictions

Understanding interactions using predictions

- The predictions are different if we ignore the interaction term:

```
p.lpm.ignore <- as.numeric(c1[1] + c1[2]*25 + c1[3])
p.logit.ignore <- invlogit(as.numeric(c2[1] + c2[2]*25 + c2[3]))
print(p.lpm.ignore)
```

```
## [1] 0.5855933
```

```
print(p.logit.ignore)
```

```
## [1] 0.6184885
```

Predictions

Understanding interactions using predictions

- We could also make the same predictions for native women, holding age constant.
- The equation is simplified since the main effect and interaction effect are now zero:

```
p.lpm2 <- as.numeric(c1[1] + c1[2]*25)
p.logit2 <- invlogit(as.numeric(c2[1] + c2[2]*25))
print(p.lpm2)
```

```
## [1] 0.2483897
```

```
print(p.logit2)
```

```
## [1] 0.2437525
```

- Despite the negative interaction, the main effect of nativity implies that a village native will be less likely to belong to an SHG, holding age constant.

Predictions

Using the `predictions` function

- We can do this systematically by creating a new object containing every combination of predictors. The `predictions` function can then be used to obtain predicted values from the model.

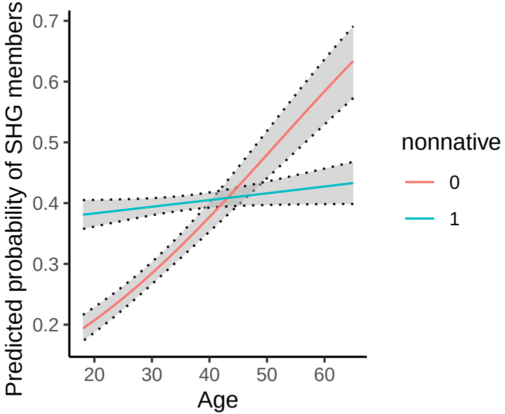
```
ages <- 18:65
nativity <- 0:1
new <- expand.grid(list("age" = ages, "nonnative" = nativity))

preds <- predictions(logistic, newdata = new)
preds %>% select(estimate, age, nonnative) %>%
    head(5) %>% kable()
```

estimate	age	nonnative
0.194	18	0
0.200	19	0
0.207	20	0
0.214	21	0
0.221	22	0

Predictions

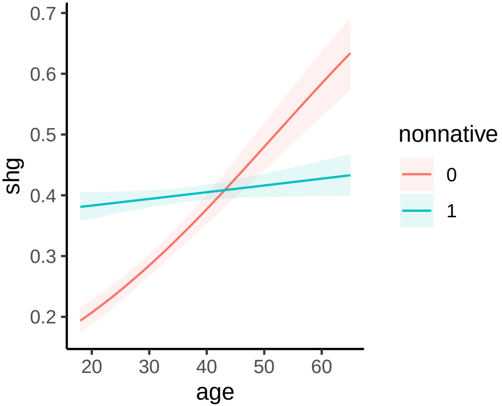
Plotting the results[2]



[2]Standard errors are calculated using an approach known as the delta method. See this post for further details.

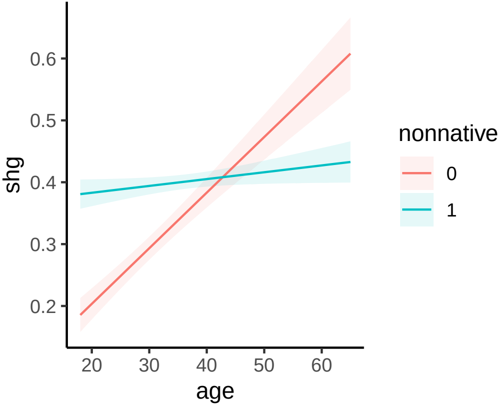
Predictions

We can directly obtain these results by using the `plot_predictions` function.



Predictions

The LPM shows a similar pattern but the predictions are constrained to be linear.



Predictions

Improving the model

- The previous model suggests differences in relationship by nativity and age:
  - For natives, there is a strong positive relationship between age and SHG membership.
  - For nonnatives, there is little evidence of such a relationship.
- Although there are age differences, these patterns seem remarkably strong.
- Let's add a squared term to account for non-linear effects of age.

Predictions

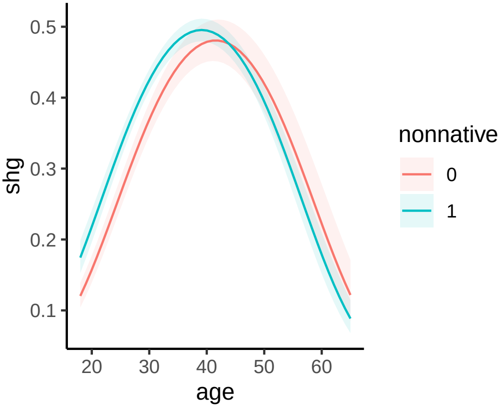
Improving the model

	Logistic 1	Logistic 2
(Intercept)	-2.184***	-6.030***
	(0.130)	(0.261)
age	0.042***	0.287***
	(0.004)	(0.014)
nonnative	1.615***	0.737***
	(0.159)	(0.181)
age $\times$ nonnative	-0.037***	-0.017***
	(0.004)	(0.005)
I(age^2)		-0.003***
		(0.000)
Num.Obs.	8976	8976
Log.Lik.	-5818.085	-5638.019
F	64.991	121.010

Note: ^^ * $p < 0.05$, ** $p < 0.01$, *** $p < 0.001$

Predictions

Making new predictions



Marginal effects

Predictions versus marginal effects

- Predictions and associated plots allow us to observe differences on the outcome scale (in this case probabilities) across different values of the data.
- But what if we want to make statements about the overall effect of a predictor?
  - What is the average effect of age?
  - How does the effect of age vary as a function of other covariates?
- Like polynomial regression, it is difficult to determine this by examining coefficients or plotting predictions.

Marginal effects

Definitions

- A **marginal effect** is the relationship between change in single predictor and the dependent variable while *holding other variables constant*.
- "Marginal effects are partial derivatives of the regression equation with respect to each variable in the model for each unit in the data."[3]
- Recall that standard OLS coefficients can be intepreted as marginal effects, but this is no longer true with logistic regression.
- In a logistic regression, the effects of predictors can be non-linear, such that the effect of x on y will vary according to the level of z.

[3] Leeper 2021. See the vignette for the margins package.

Marginal effects

Calculation

- All analyses in this lecture use the `marginaleffects` package, which extends the functionality of `margins` and works for both frequentist and Bayesian models.[4]

[4] Read the documentation provided here for further information.

Marginal effects

Calculation

Observe how the `slopes` function returns $N * k$ rows, where N is the number of observations in the dataset and k is the number of *unique* predictors.

```
ME <- slopes(logistic2)
dim(ME)
```

```
## [1] 17952    15
```

```
dim(data)[1]*2
```

```
## [1] 17952
```

Marginal effects

Interpretation

This table shows the marginal effects for age and nonnative for the first two respondents.

```
ME %>% filter(rowid <= 2) %>% arrange(rowid) %>%
    select(rowid, term, contrast, estimate, std.error, shg, age, nonnative) %>%
    kable()
```

rowid	term	contrast	estimate	std.error	shg	age	nonnative
1	age	dY/dX	0.020	0.001	0	27	1
1	nonnative	1 - 0	0.063	0.014	0	27	1
2	age	dY/dX	0.022	0.001	1	24	1
2	nonnative	1 - 0	0.066	0.015	1	24	1

Marginal effects

Marginal effects at specified values

- Marginal effects are better understood by contextualizing them at relevant values of the data.
- Like the example above, we may want to calculate the marginal effect of a predictor at specific values of other covariates.
- e.g. How does the effect of age vary by nativity and age?

Marginal effects

Marginal effects at specified values

```
ME.n <- slopes(logistic,
               newdata = datagrid(nonnative = c(0,1),
                                  age = c(25)))
ME.n %>% filter(term == "age") %>%
    select(estimate, std.error, nonnative, age) %>%
    kable()
```

estimate	std.error	nonnative	age
0.008	0.001	0	25
0.001	0.001	1	25

Marginal effects

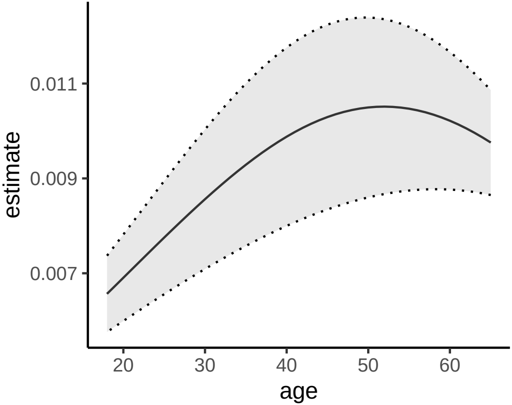
Marginal effects at specified values

```
ME.a <- slopes(logistic,
               newdata = datagrid(nonnative = 0,
                                  age = 18:65))
ME.a %>% filter(term == "age") %>%
    select(estimate, std.error, age)  %>%
    filter(age %in% c(18,25,35,45,55,65)) %>%
    head() %>% kable()
```

estimate	std.error	age
0.007	0.000	18
0.008	0.001	25
0.009	0.001	35
0.010	0.001	45
0.010	0.001	55
0.010	0.001	65

Marginal effects

Marginal effects at specified values

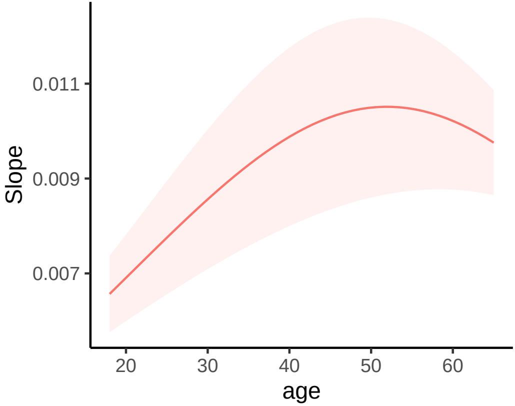


Note the non-linear relationship occurs even thought the inputs to the model are linear. This is because the logistic regression creates a non-linear mapping of the linear model.

Marginal effects

Plotting conditional marginal effects using `plot_slopes`

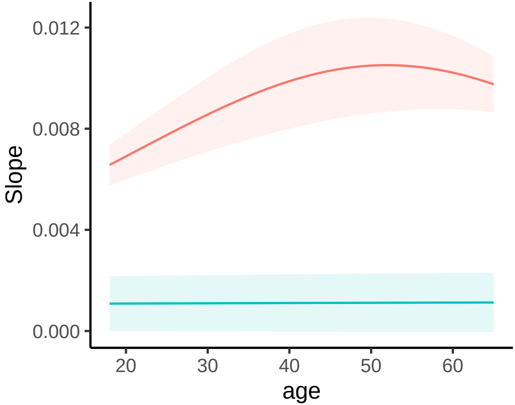
```
plot_slopes(logistic, variables = "age",
            condition = list("age", "nonnative" = 0)) +
    theme_classic() + theme(legend.position = "none")
```



Marginal effects

Plotting conditional marginal effects using `plot_slopes`

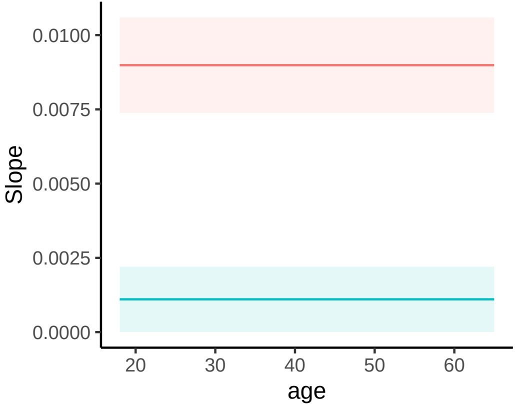
```
plot_slopes(logistic, variables = "age",
            condition = list("age", "nonnative")) +
    theme_classic() + theme(legend.position = "none")
```



Marginal effects

Comparision with the LPM

```
plot_slopes(lpm, variables = "age",
            condition = list("age", "nonnative")) +
    theme_classic() + theme(legend.position = "none")
```

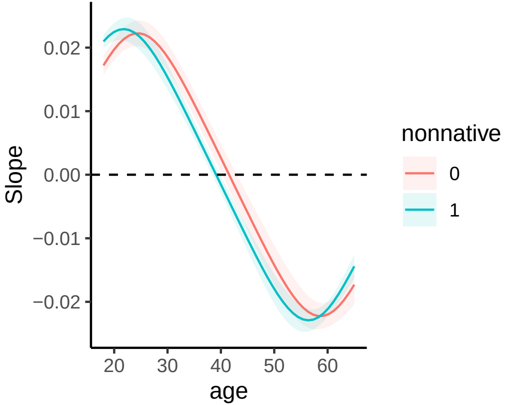


Marginal effects

Plotting conditional marginal effects using `plot_slopes`

The relationship changes substantially when we add age^2.

```
plot_slopes(logistic2, variables = "age", condition = c("age", "nonnative")) +
    theme_classic() + geom_hline(yintercept = 0, linetype = "dashed")
```



Marginal effects

Marginal effects at means

- A common approach is to assess the **marginal effects at means (MEM)**, examining the marginal effect of change in a predictor while holding other covariates at their average values.
- This can be convenient if we don't have any clear reasons for selecting particular values to examine.

Marginal effects

Marginal effects at means

We can get the marginal effects at means by specifying `newdata = "mean"`. Does this approach make sense in this case?

```
slopes(logistic2, newdata = "mean") %>%
    select(term, estimate, age, nonnative) %>%
    kable()
```

term	estimate	age	nonnative
age	0.008	34.951	0.719
nonnative	0.037	34.951	0.719

Marginal effects

Marginal effects at means

In this case, it is more appropriate to consider the marginal effects for each value of nonnative. By default, age is now held at the mean value.

```
slopes(logistic2,
                  newdata = datagrid(nonnative = c(0,1))) %>%
                  select(term, estimate, age, nonnative) %>%
                  kable()
```

term	estimate	age	nonnative
age	0.011	34.951	0
age	0.007	34.951	1
nonnative	0.037	34.951	0
nonnative	0.037	34.951	1

Marginal effects

Average marginal effects

- Another approach involves averaging over the variation in other covariates to calculate the **average marginal effect (AME)** of a predictor.
- We can obtain this by averaging over all the observation specific marginal effects.

Marginal effects

Average marginal effects

We can obtain the AME by taking a summary of the marginal effects table produced above (`ME`).

```
AME <- summary(ME)
AME %>% select(term, estimate, std.error) %>% kable()
```

term	estimate	std.error
age	0.006	0.000
nonnative	0.029	0.012

Marginal effects

Average marginal effects

This quantity can also be directly computed using `avg_slopes`.

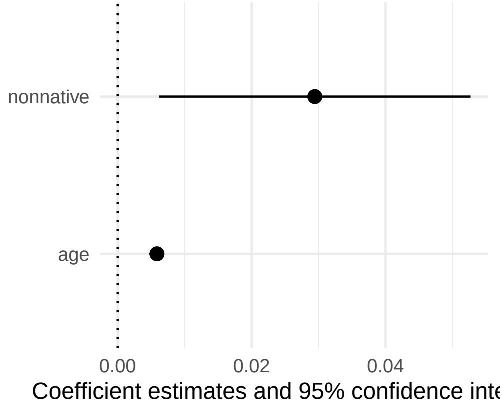
```
avg_slopes(logistic2) %>%
                select(term, estimate, std.error) %>%
                kable()
```

term	estimate	std.error
age	0.006	0.000
nonnative	0.029	0.012

Marginal effects

We can plot of the marginal effects and confidence intervals by calling `modelplot` on the marginal effects table.

```
modelplot(ME) + geom_vline(xintercept = 0, linetype = "dotted")
```



Marginal effects

Full specification

- These models show how the marginal effect of age is highly non-linear
- Let's add some complexity by incorporating covariates for caste and education
- I also add the village-level fixed effects to account for spatial variation

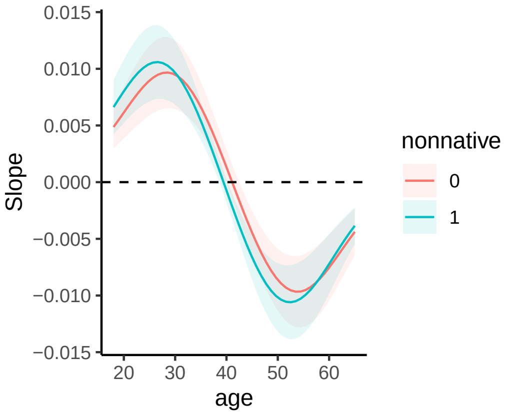
Marginal effects

	Logistic 1	Logistic 2	Logistic 3
(Intercept)	-2.184***	-6.030***	-7.688***
	(0.130)	(0.261)	(0.402)
age	0.042***	0.287***	0.322***
	(0.004)	(0.014)	(0.016)
nonnative	1.615***	0.737***	0.650***
	(0.159)	(0.181)	(0.193)
age × nonnative	-0.037***	-0.017***	-0.013*
	(0.004)	(0.005)	(0.005)
I(age^2)		-0.003***	-0.004***
		(0.000)	(0.000)
castelow			0.320***
			(0.056)
educ			0.001
			(0.007)
Log.Lik.	-5818.085	-5638.019	-5049.587

Note: ^^ * $p < 0.05$, ** $p < 0.01$, *** $p < 0.001$

Marginal effects

```
plot_slopes(logistic3, variable = "age", condition = c("age", "nonnative")) +
    theme_classic() + geom_hline(yintercept = 0, linetype = "dashed")
```



Marginal effects

Bayesian estimation

- The same approaches apply to Bayesian models. The only difference is that the uncertainty in the posterior distribution must be incorporated into the calculation of the marginal effects.
- Fortunately for us, the `marginaleffects` package can handle models estimated using `rstanarm`.

Marginal effects

Bayesian estimation

```
bayes <- stan_glm(shg ~ age + I(age^2) + nonnative + age:nonnative,
                  data = data, family = binomial(),
                  chains = 1, refresh = 0)
```

Marginal effects

Bayesian estimation

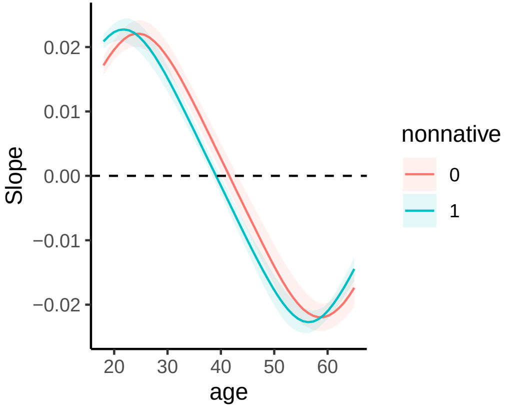
The AMEs are close to those obtained from the maximum likelihood model.

```
summary(slopes(bayes)) %>%
    kable()
```

term	contrast	estimate	conf.low	conf.high
age	mean(dY/dX)	0.006	0.005	0.007
nonnative	mean(1) - mean(0)	0.030	0.005	0.053

Marginal effects

We can see similar relationships using the same `plot_slopes` specification as above.



Marginal effects

Improving the model?

Let's fit something even more complex using `stan_glm`.

```
bayes.2 <- stan_glm(shg ~ age + I(age^2) + nonnative +
                        caste + nonnative:caste +
                        age:caste + age:nonnative +
                        educ + village,
                    data = data, family = binomial(),
                    chains = 1,  refresh = 0)
```

Marginal effects

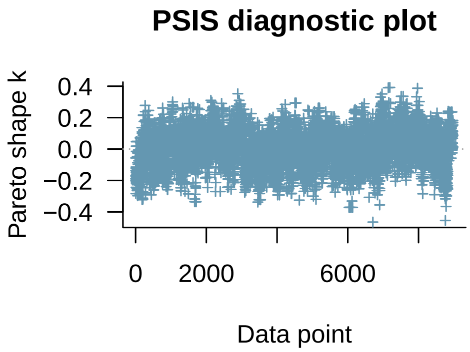
Comparing the held-out likelihood scores using LOO-CV

```
l1 <- loo(bayes)
l2 <- loo(bayes.2)
loo_compare(l1,l2)
```

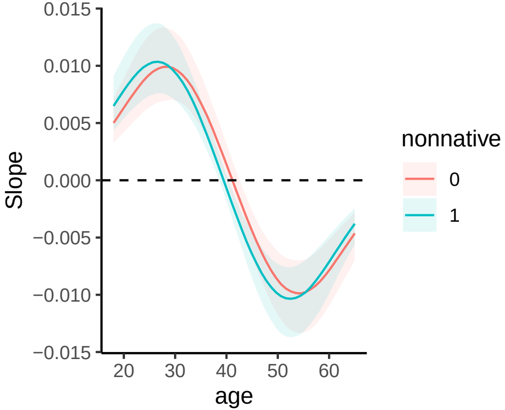
```
##         elpd_diff se_diff
## bayes.2    0.0       0.0
## bayes   -511.8      32.7
```

Marginal effects

```
plot(l2)
```

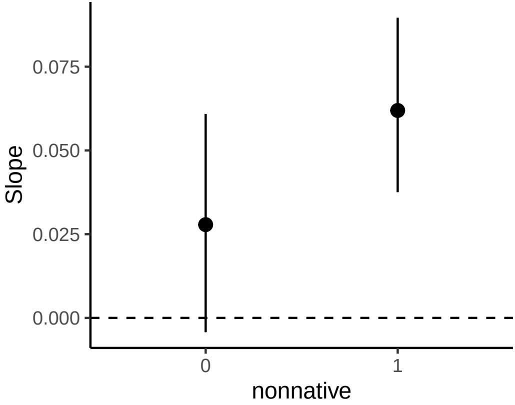


Marginal effects



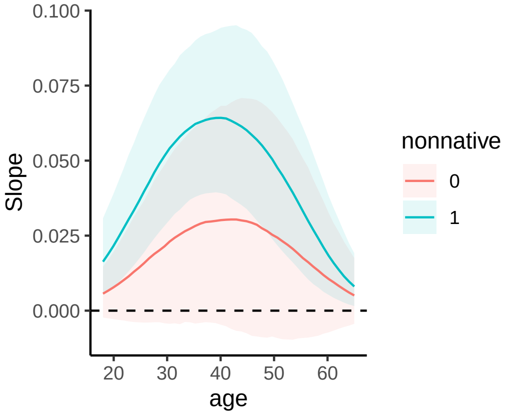
Marginal effects

```
plot_slopes(bayes.2, variable = "caste", condition = c("nonnative")) +
    theme_classic() + geom_hline(yintercept = 0, linetype = "dashed")
```



Marginal effects

```
plot_slopes(bayes.2, variable = "caste", condition = c("age", "nonnative")) +
    theme_classic() + geom_hline(yintercept = 0, linetype = "dashed")
```



Summary

- Logistic regression models (and other GLMs) can be challenging to interpret, particularly when we add interaction terms.
- By making predictions, we can observe variation in outcomes across different values and interpret results on the outcome scale.
- Marginal effects allow us to isolate the effect of individual variables, akin to the way we interpret OLS results, and to assess relationships between predictions.
- In both cases, visualizations improve our understanding of the relationships between variables compared to regression tables alone.

Next week

- Count outcomes
- Poisson regression
- Negative-binomial regression
- And zero-inflated variants

Lab

- Logistic regression, interactions, and marginal effects.