

Lesson 4:

Multiple Regression

This Lesson's Goals

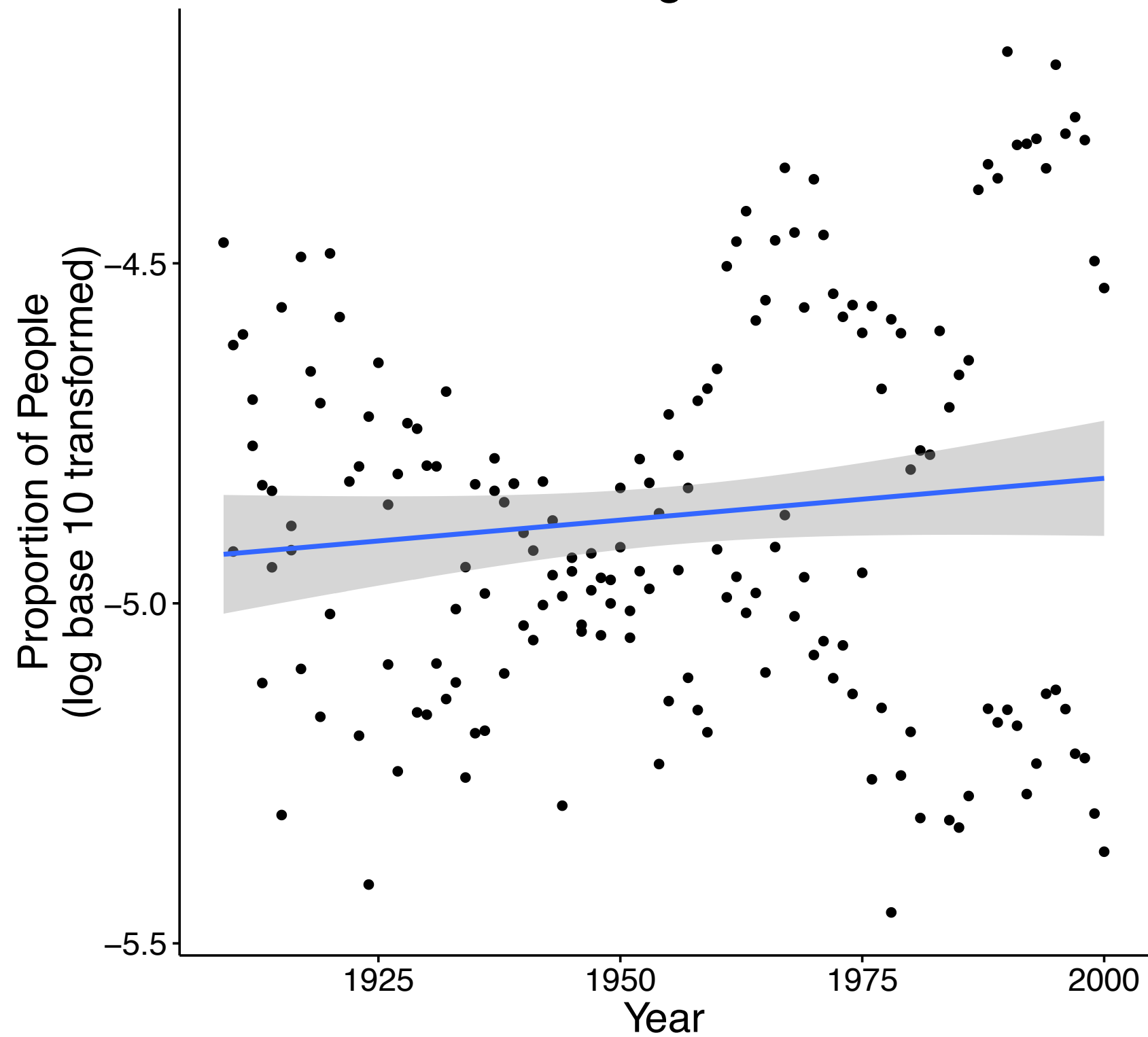
Learn about multiple regression

Make a figure for data for a multiple regression

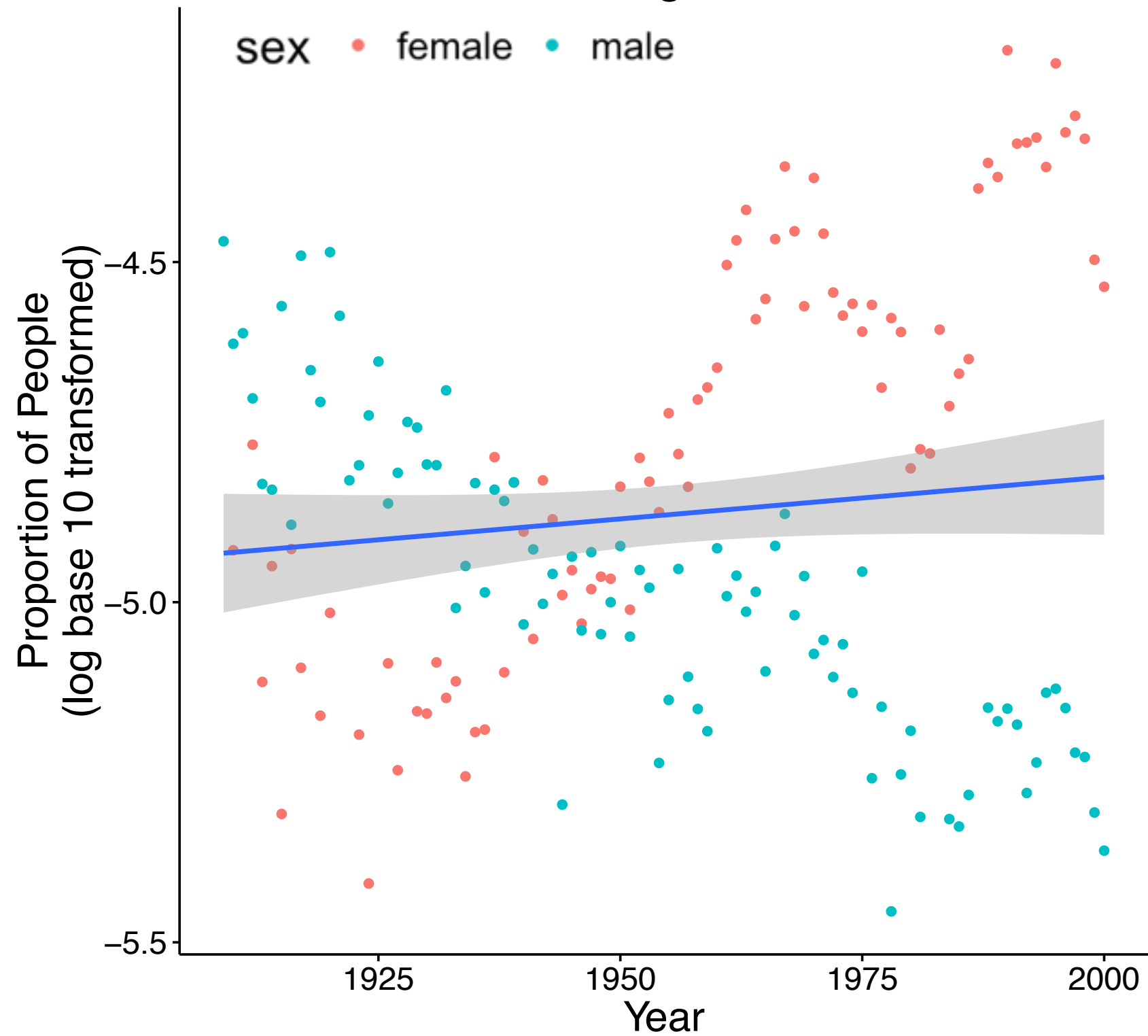
Do a multiple regression in R

Summarise results in an R Markdown document

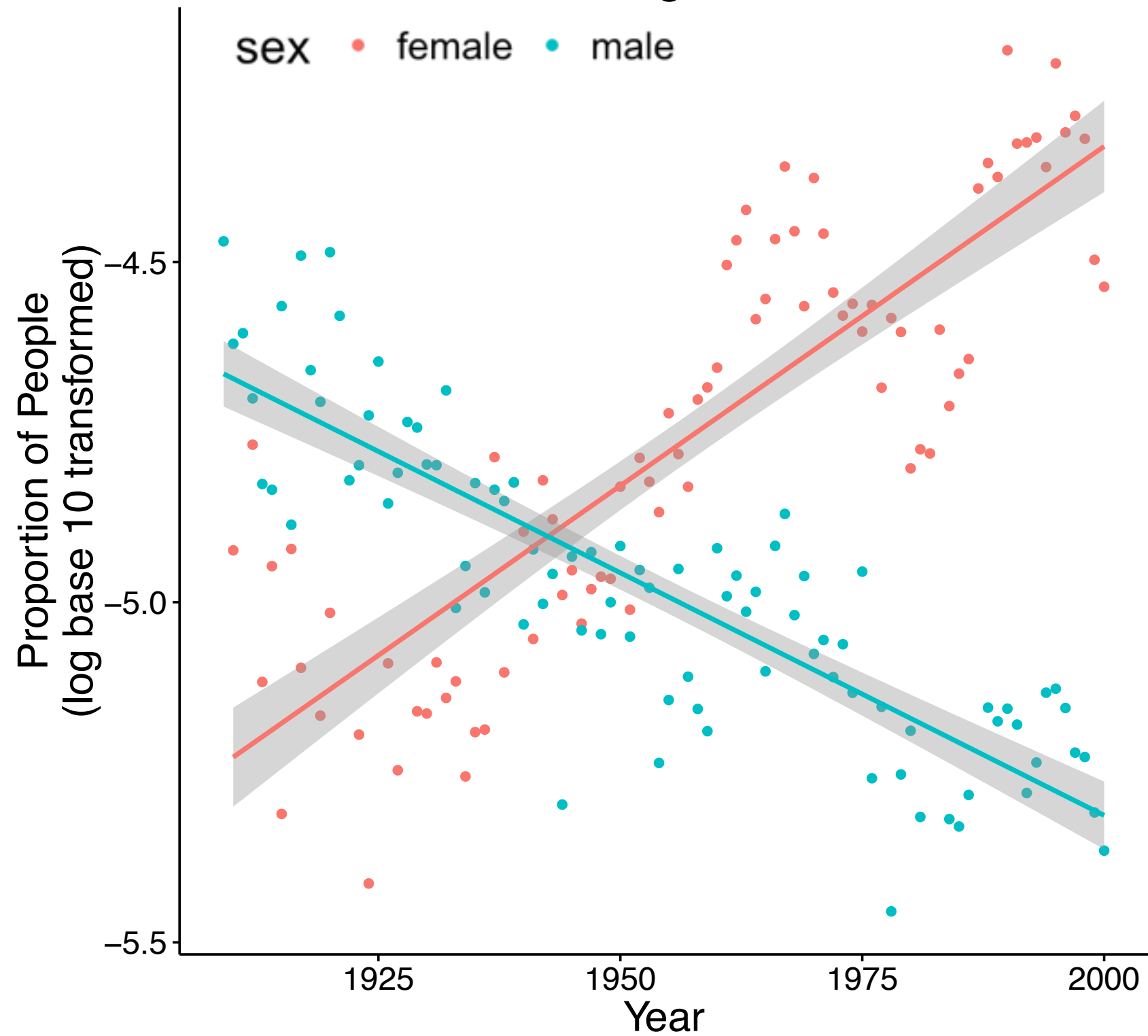
Proportion of People with the Name 'Page' Over Time



Proportion of People with the Name 'Page' Over Time



Proportion of People with the Name 'Page' Over Time



Math (Part 1)

$$y_i = a + bx_i + e_i$$

**What if you have two
variables?**

$$y_i = a + bx_i + e_i$$

$$y_i = a + b_1X_{1i} + b_2X_{2i} + e_i$$

new
variable



$$y_i = a + b_1 x_{1i} + b_2 x_{2i} + e_i$$

y_i = specific y value

a = intercept

b_1 = slope of first variable

x_{1i} = specific x value for
first variable

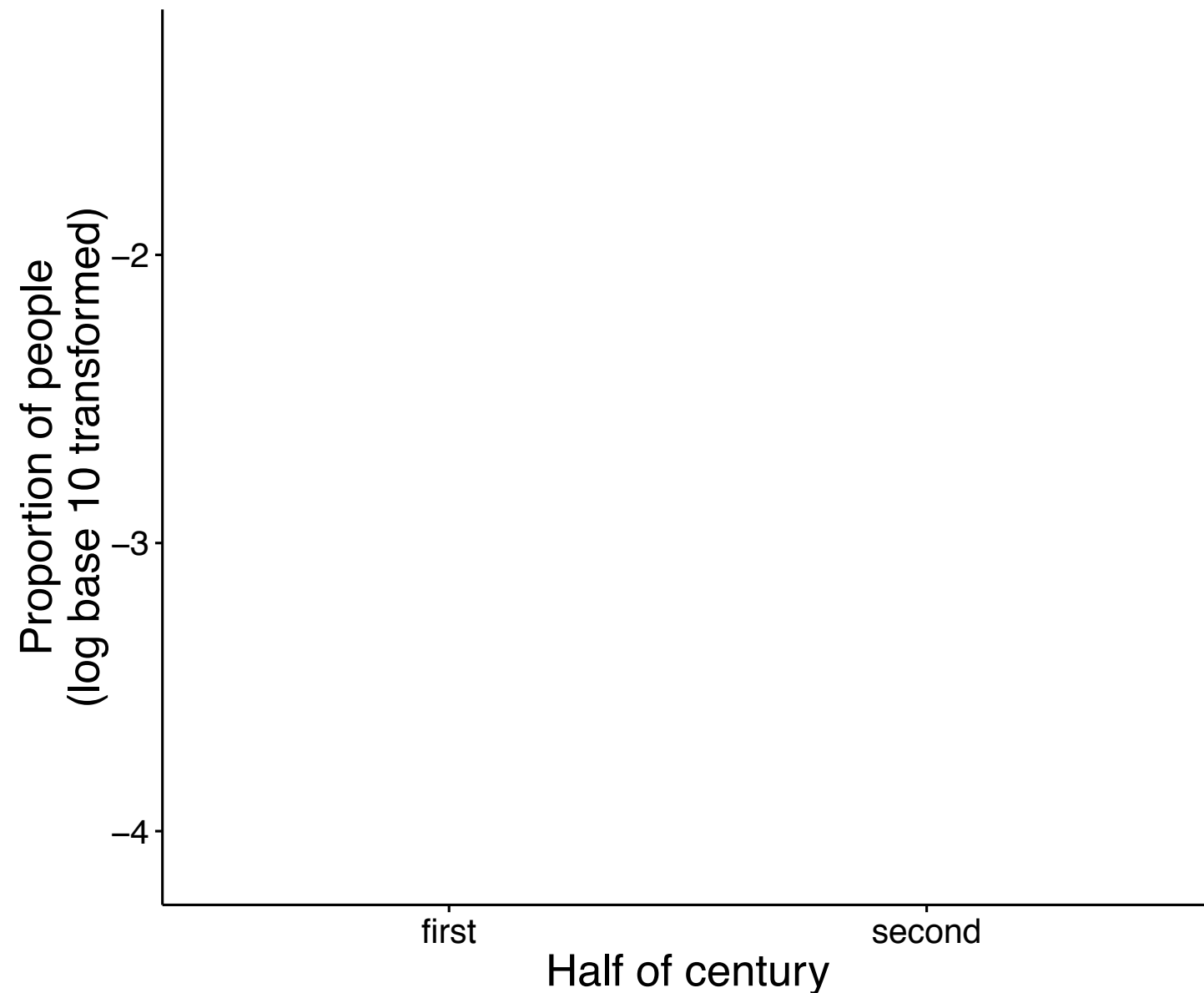
b_2 = slope of second
variable

x_{2i} = specific x value for
second variable

e_i = random variance or the
residual

$$y_i = a + b_1X_{1i} + b_2X_{2i} + e_i$$

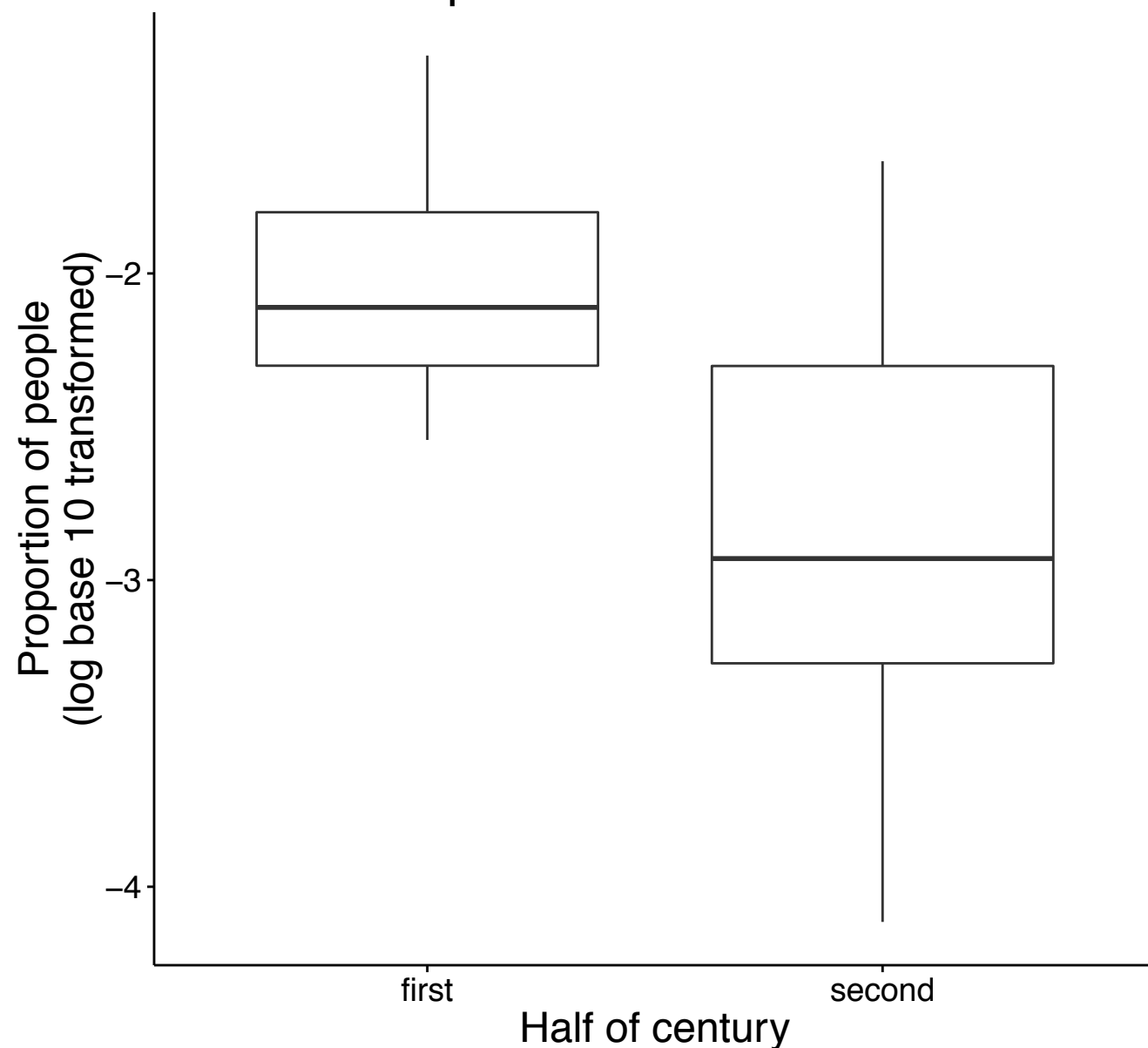
Proportion of People with
Popular Names for 1901



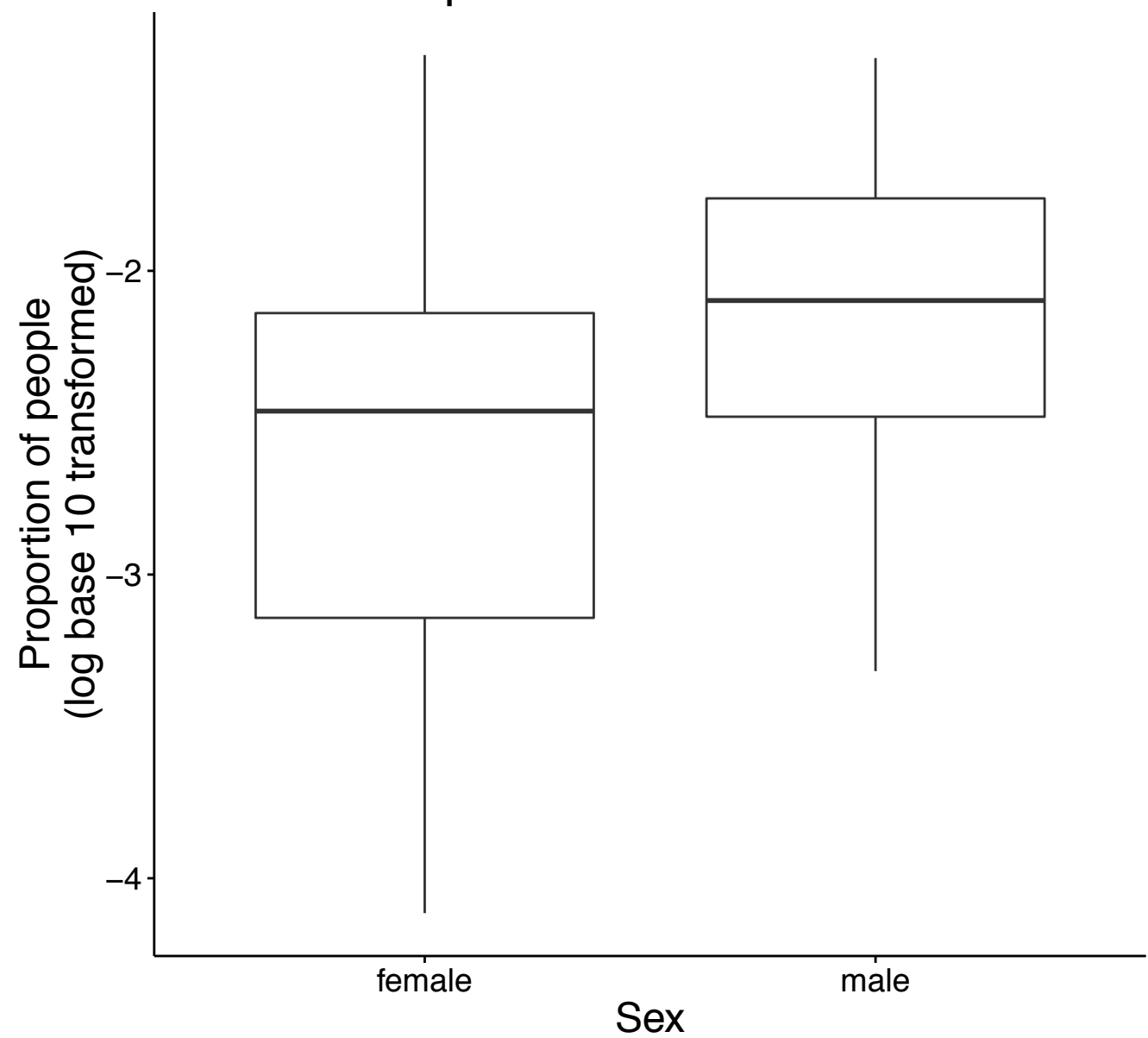
$$y_i = a + b_1 X_{1i} + b_2 X_{2i} + e_i$$

$$y_i = a + b_1 X_{1i} + e_i \quad y_i = a + b_2 X_{2i} + e_i$$

Proportion of People with
Popular Names for 1901

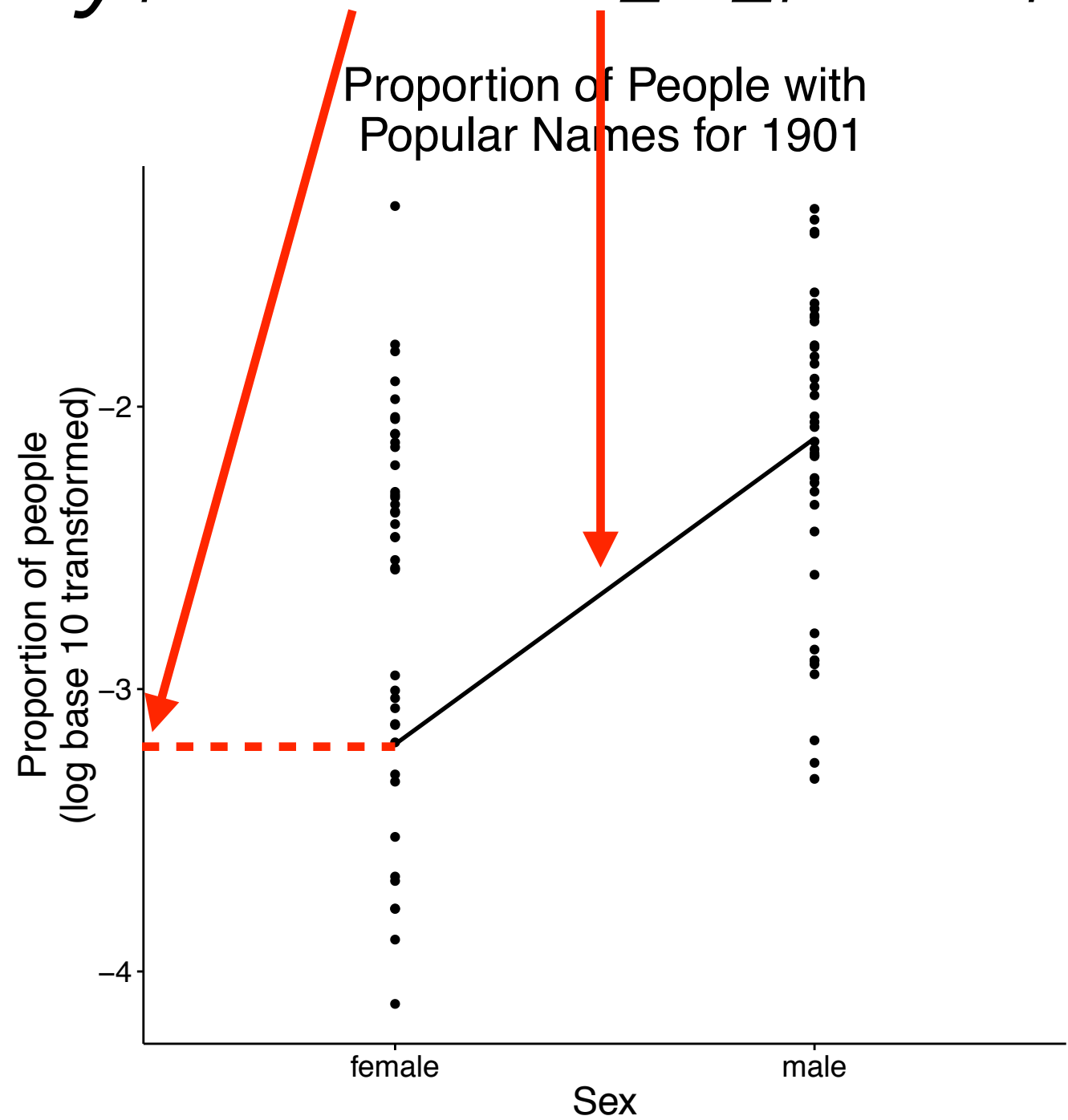
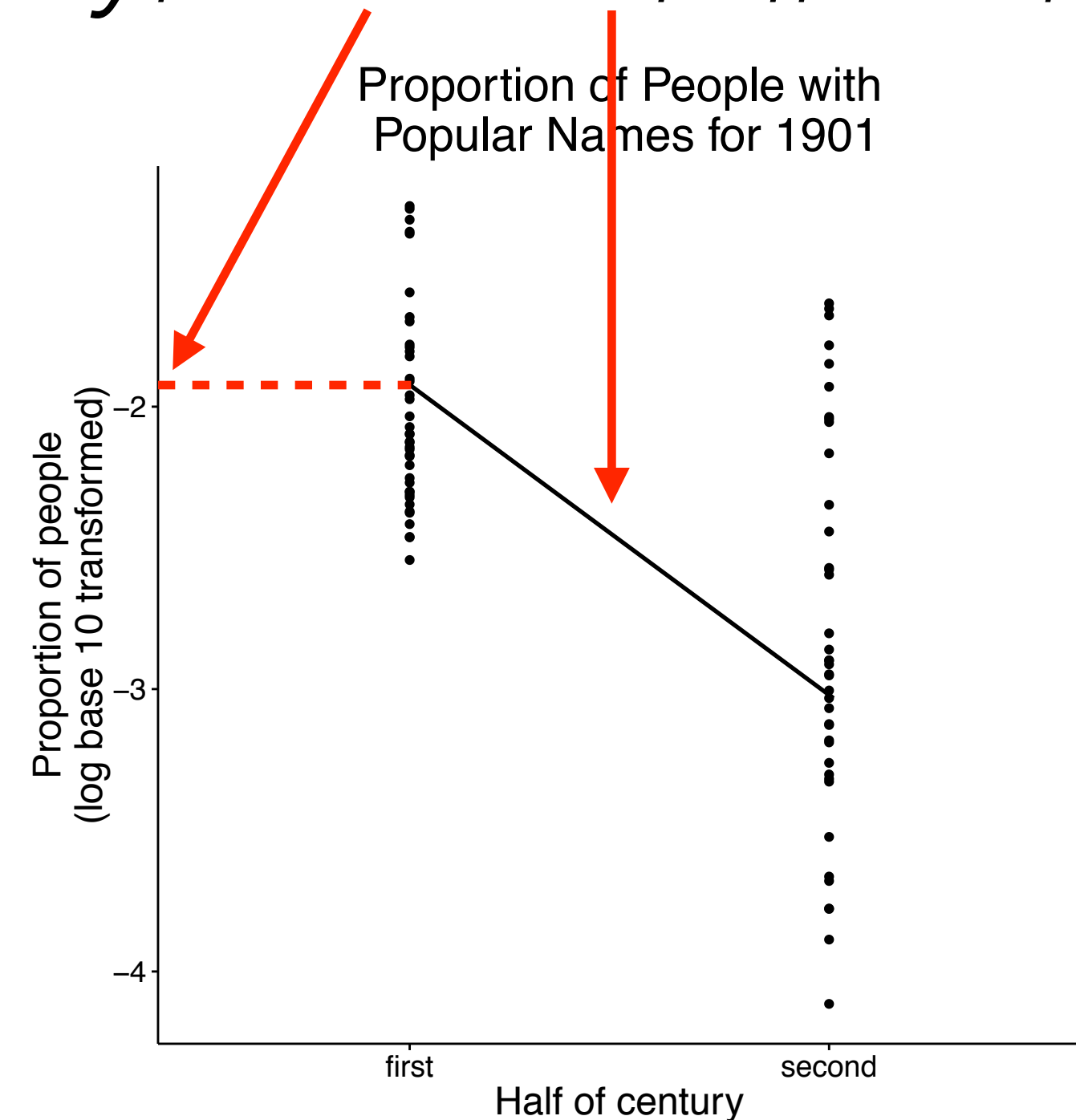


Proportion of People with
Popular Names for 1901



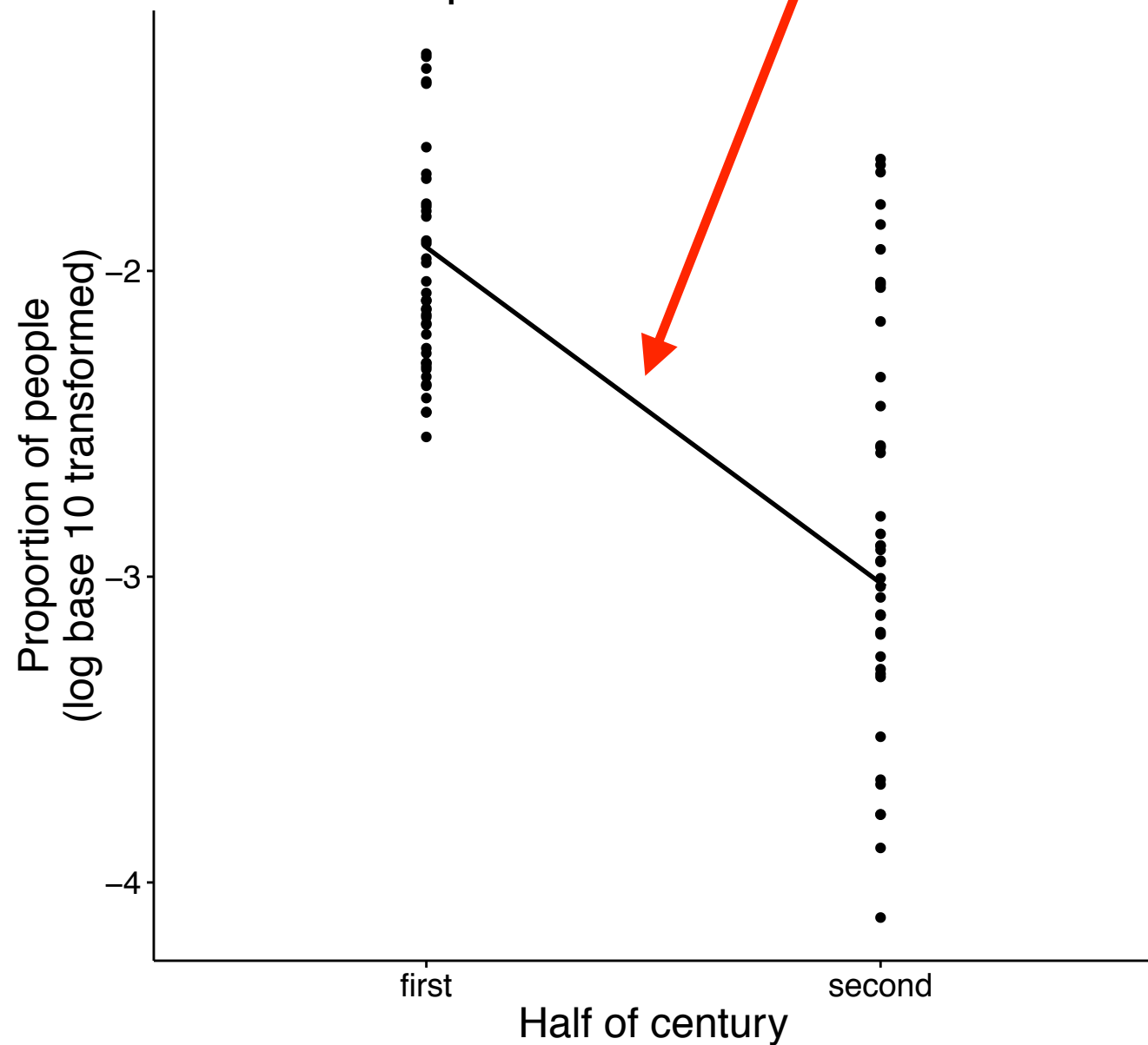
$$y_i = a + b_1X_{1i} + b_2X_{2i} + e_i$$

$$y_i = a + b_1X_{1i} + e_i \quad y_i = a + b_2X_{2i} + e_i$$

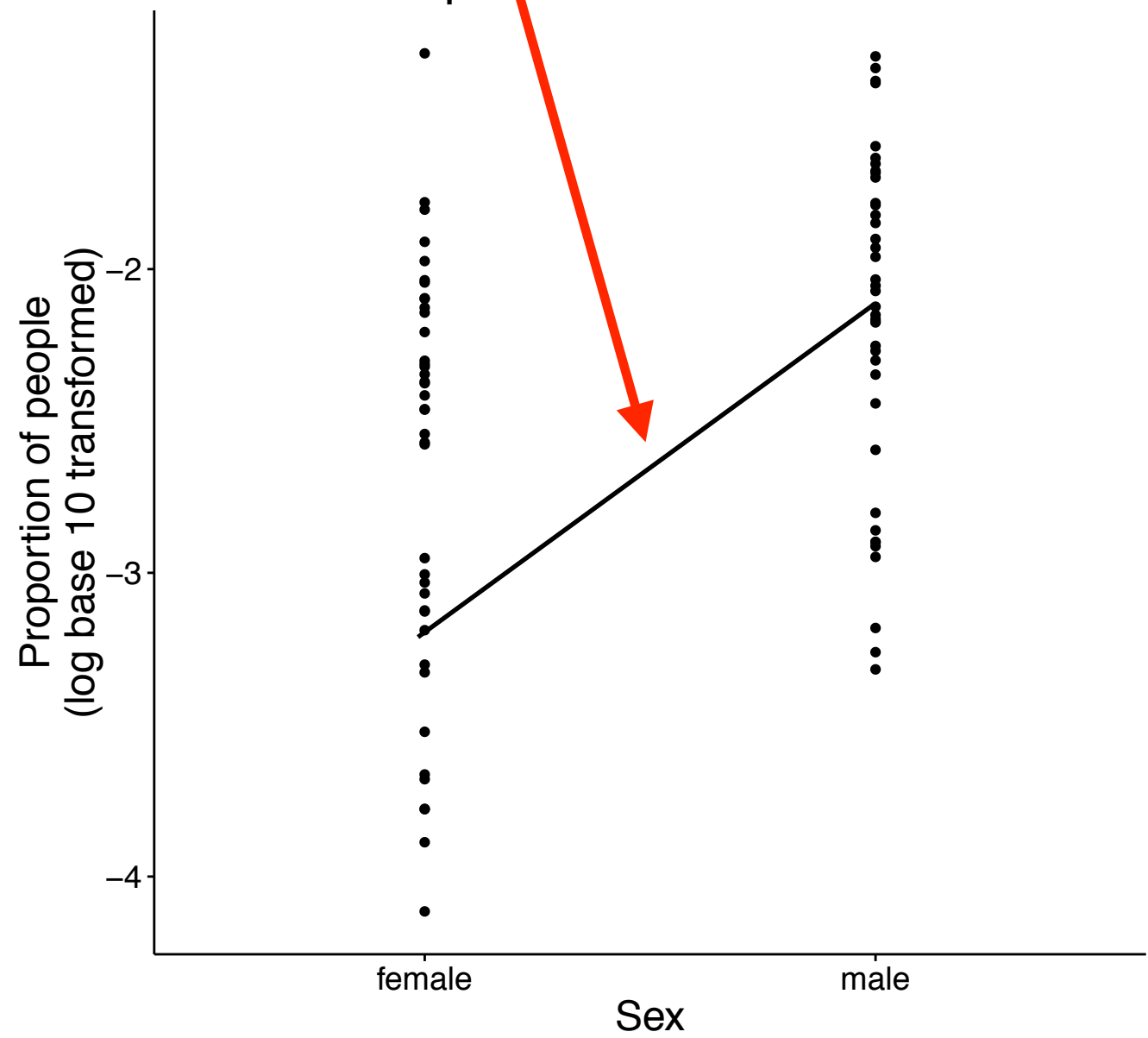


$$y_i = a + b_1 X_{1i} + b_2 X_{2i} + e_i$$

Proportion of People with Popular Names for 1901



Proportion of People with Popular Names for 1901



What is '*a*'?

complicated

“In conclusion, when you fit an additive model..., the parameters are the difference of the mean per category (of only one factor) and the intercept is the **estimated** value of the response variable for the first modalities of each factor **under the assumption of additivity.**”

*Stack Exchange, gui11aume
emphasis original*

R Code (Part 1)

$$y_i = a + b_1 X_{1i} + b_2 X_{2i} + e_i$$

`lm(prop_log10_mean ~ century_half + sex)`

Call:

`lm(formula = prop_log10_mean ~ century_half + sex, data = data_names)`

Residuals:

Min	1Q	Median	3Q	Max
-1.03372	-0.32975	-0.04454	0.31720	1.04439

Coefficients:

	Estimate	Std. Error	t value	Pr(> t)
(Intercept)	-2.2748	0.09182	-24.776	< 2e-16 ***
century_halfsecond	-0.80612	0.10602	-7.603	5.88e-11 ***
sexmale	0.51263	0.10602	4.835	6.66e-06 ***

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

Residual standard error: 0.4741 on 77 degrees of freedom

Multiple R-squared: 0.5132, Adjusted R-squared: 0.5006

F-statistic: 40.6 on 2 and 77 DF, p-value: 9.148e-13

`> head(resid(popnames.lm))`

1	2	3
-0.38859173	-0.32953459	0.17713693
4	5	6
-0.04286621	0.17908427	0.51000543

simple regression with one variable (century_half)

```
lm(prop_log10_mean ~ century_half)
```

	Estimate	Std. Error	t value	Pr(> t)	
(Intercept)	-2.01854	0.08504	-23.735	< 2e-16	***
century_halfsecond	-0.80612	0.12027	-6.703	2.88e-09	***

simple regression with one variable (sex)

```
lm(prop_log10_mean ~ sex)
```

	Estimate	Std. Error	t value	Pr(> t)	
(Intercept)	-2.67791	0.09856	-27.171	< 2e-16	***
sexmale	0.51263	0.13938	3.678	0.00043	***

multiple regression with two variables

```
lm(prop_log10_mean ~ century_half + sex)
```

	Estimate	Std. Error	t value	Pr(> t)	
(Intercept)	-2.27486	0.09182	-24.776	< 2e-16	***
century_halfsecond	-0.80612	0.10602	-7.603	5.88e-11	***
sexmale	0.51263	0.10602	4.835	6.66e-06	***

I thought the whole point of this was
that there was an interaction.

This model doesn't account for that.

**Let's run a
multiple regression with an interaction.**

Math (Part 2)

$$y_i = a + b_1X_{1i} + b_2X_{2i} + e_i$$

$$y_i = a + b_1X_{1i} \times b_2X_{2i} + e_i$$

$$y_i = a + b_1X_{1i} + b_2X_{2i} +$$

$$\boxed{b_3}X_{1i}X_{2i} + e_i$$

new
coefficient



$$y_i = a + b_1 x_{1i} + b_2 x_{2i} + b_3 x_{1i} x_{2i} + e_i$$

y_i = specific y value

b_2 = slope of second variable

a = intercept

x_{2i} = specific x value for second variable

b_1 = slope of first variable

x_{1i} = specific x value for first variable

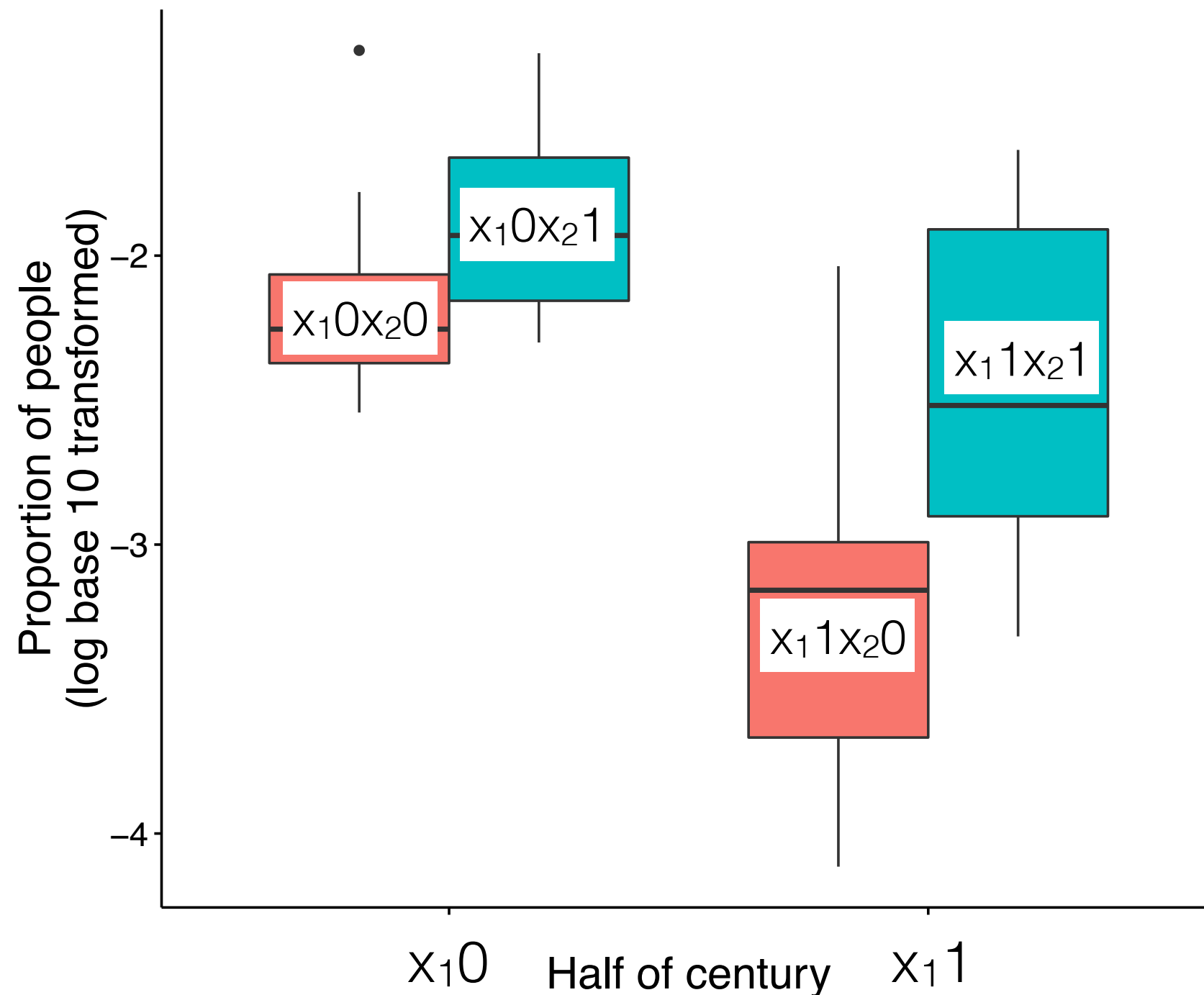
b_3 = slope of third variable (interaction)

e_i = random variance or the residual

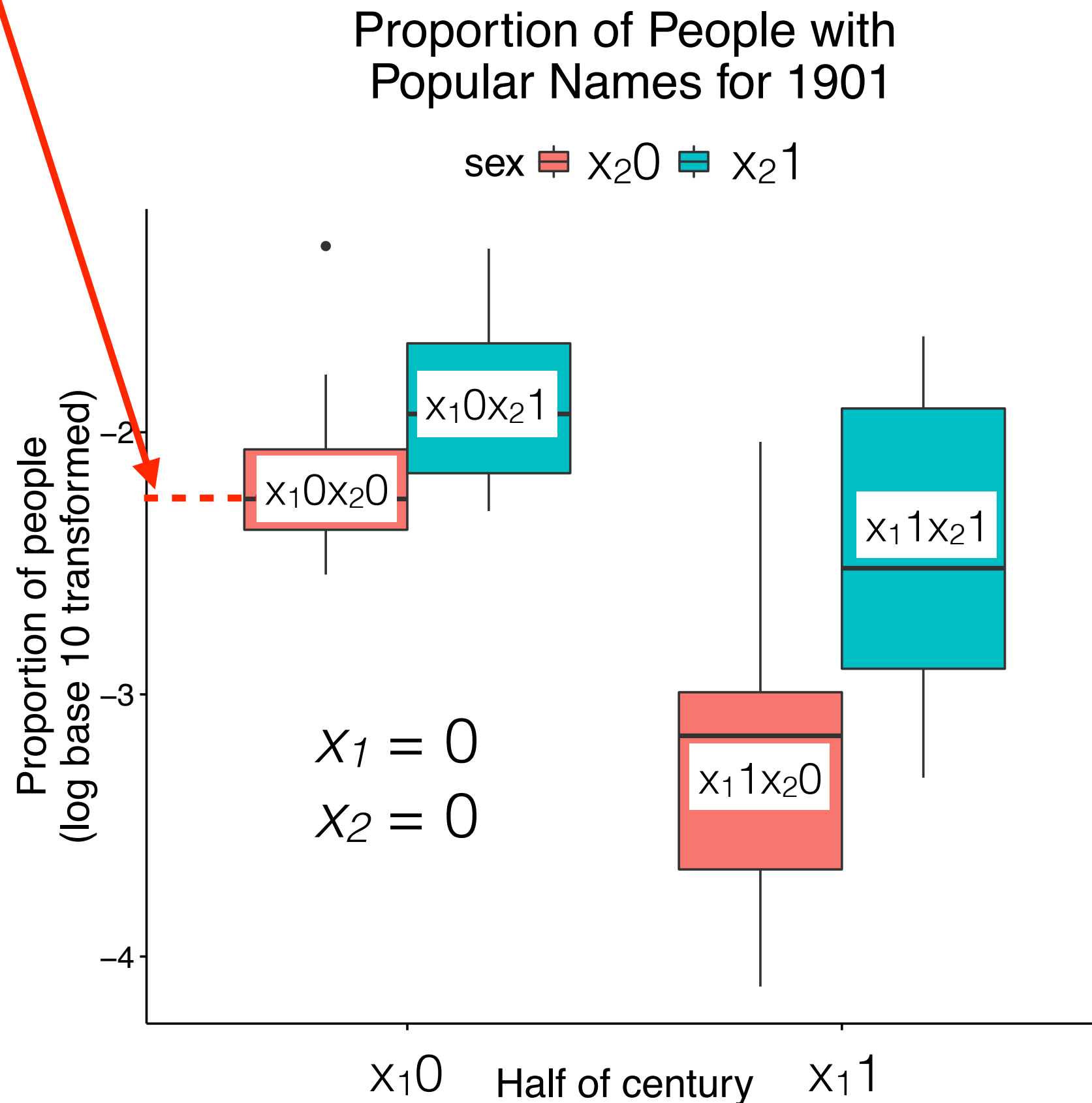
$$y_i = a + b_1x_{1i} + b_2x_{2i} + b_3x_{1i}x_{2i} + e_i$$

Proportion of People with
Popular Names for 1901

sex ■ x_{20} ■ x_{21}



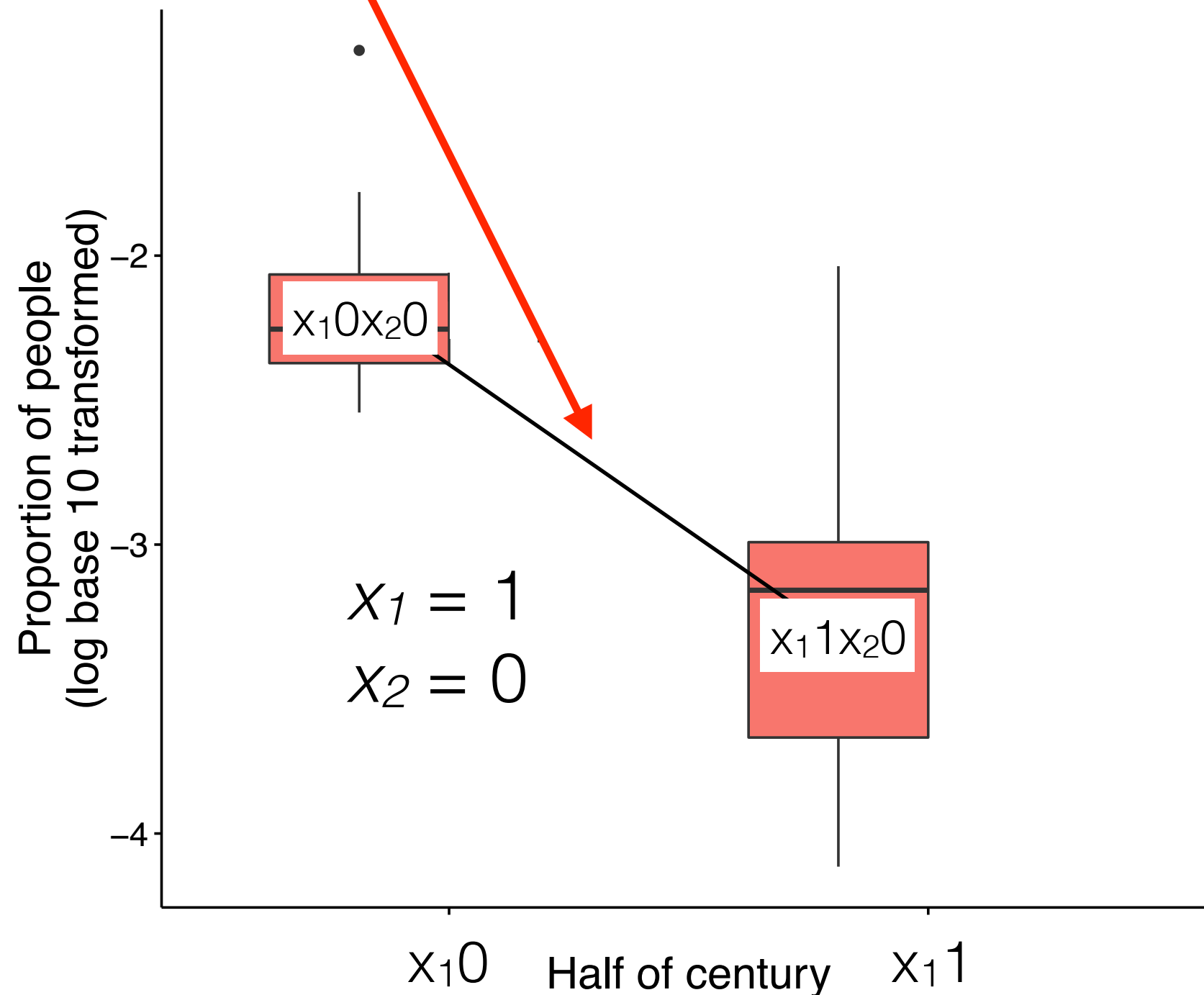
$$y_i = a + b_1x_{1i} + b_2x_{2i} + b_3x_{1i}x_{2i} + e_i$$



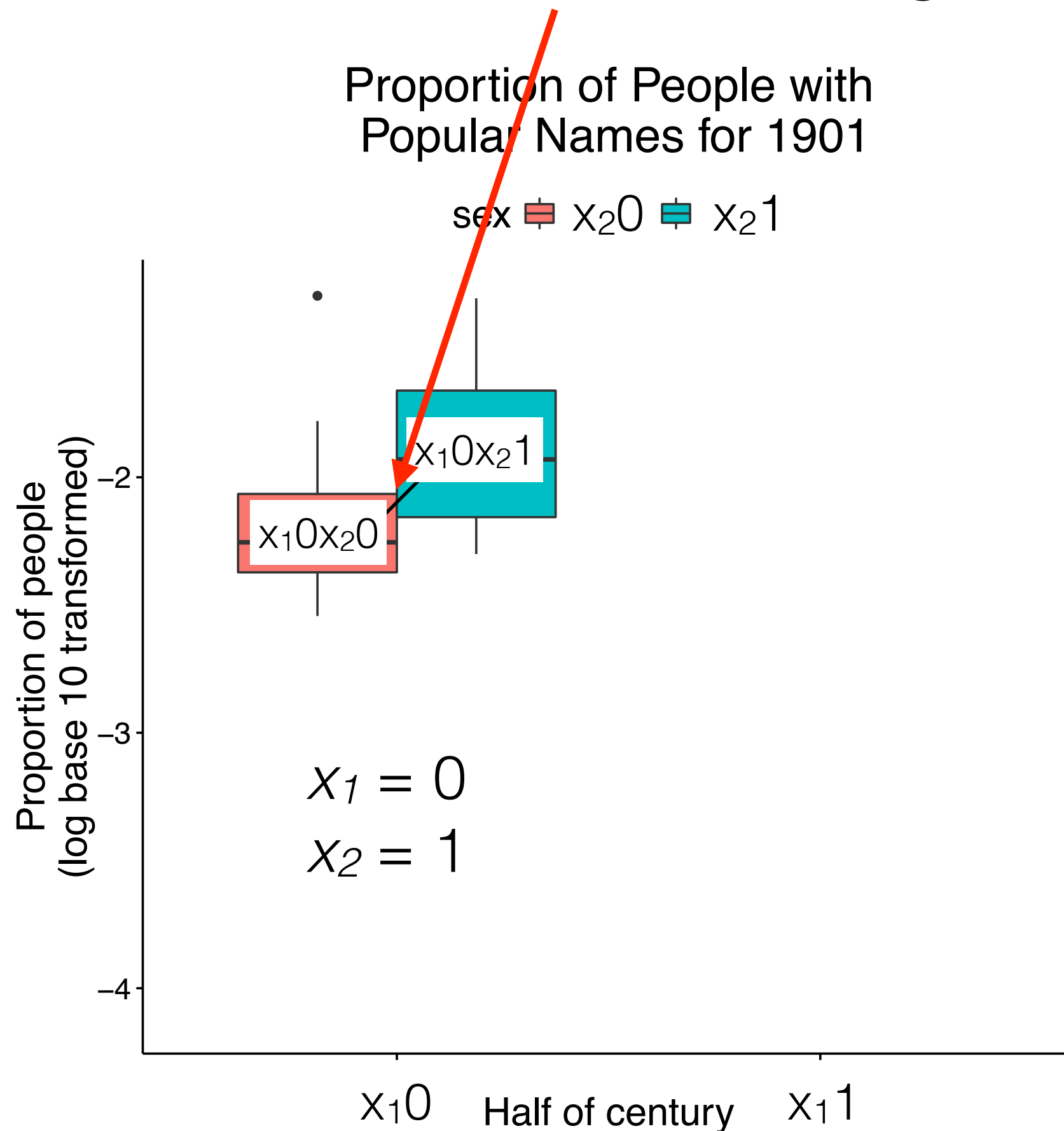
$$y_i = a + b_1x_{1i} + b_2x_{2i} + b_3x_{1i}x_{2i} + e_i$$

Proportion of People with
Popular Names for 1901

sex ■ x_{20} ■ x_{21}



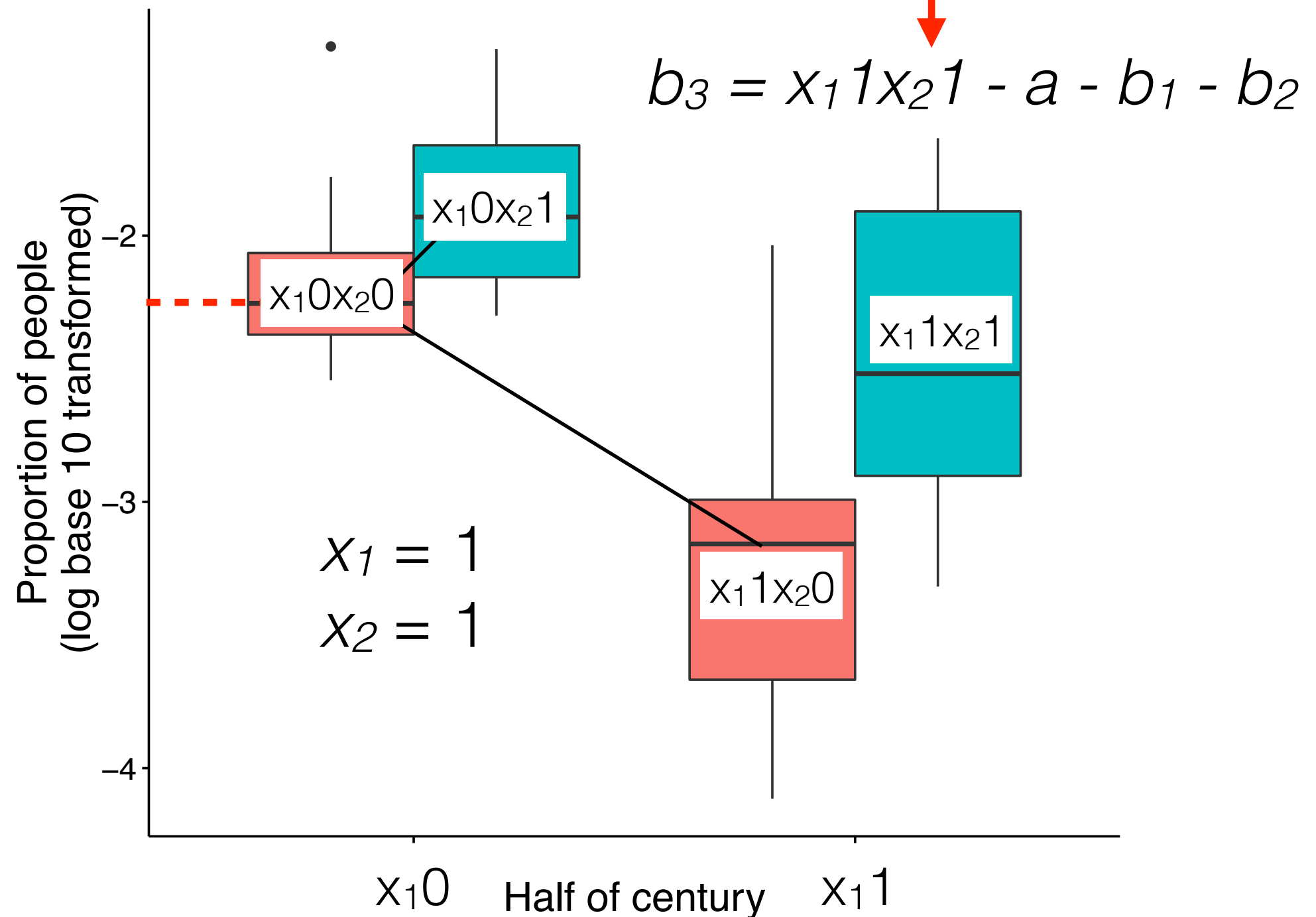
$$y_i = a + b_1x_{1i} + b_2x_{2i} + b_3x_{1i}x_{2i} + e_i$$



$$y_i = a + b_1x_{1i} + b_2x_{2i} + b_3x_{1i}x_{2i} + e_i$$

Proportion of People with
Popular Names for 1901

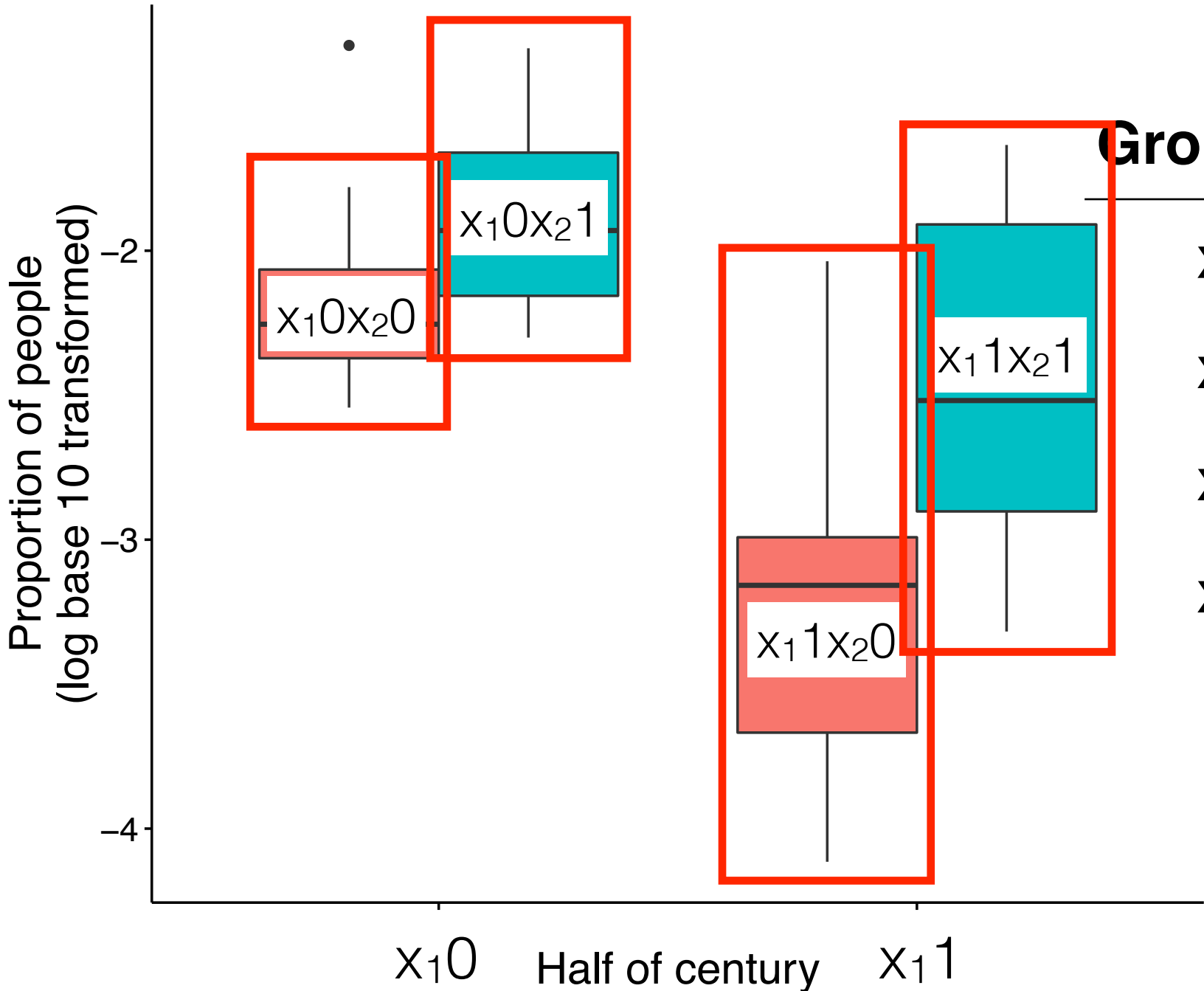
sex ■ x_{20} ■ x_{21}



$$y_i = a + b_1x_{1i} + b_2x_{2i} + b_3x_{1i}x_{2i} + e_i$$

Proportion of People with Popular Names for 1901

sex ■ x_{20} ■ x_{21}

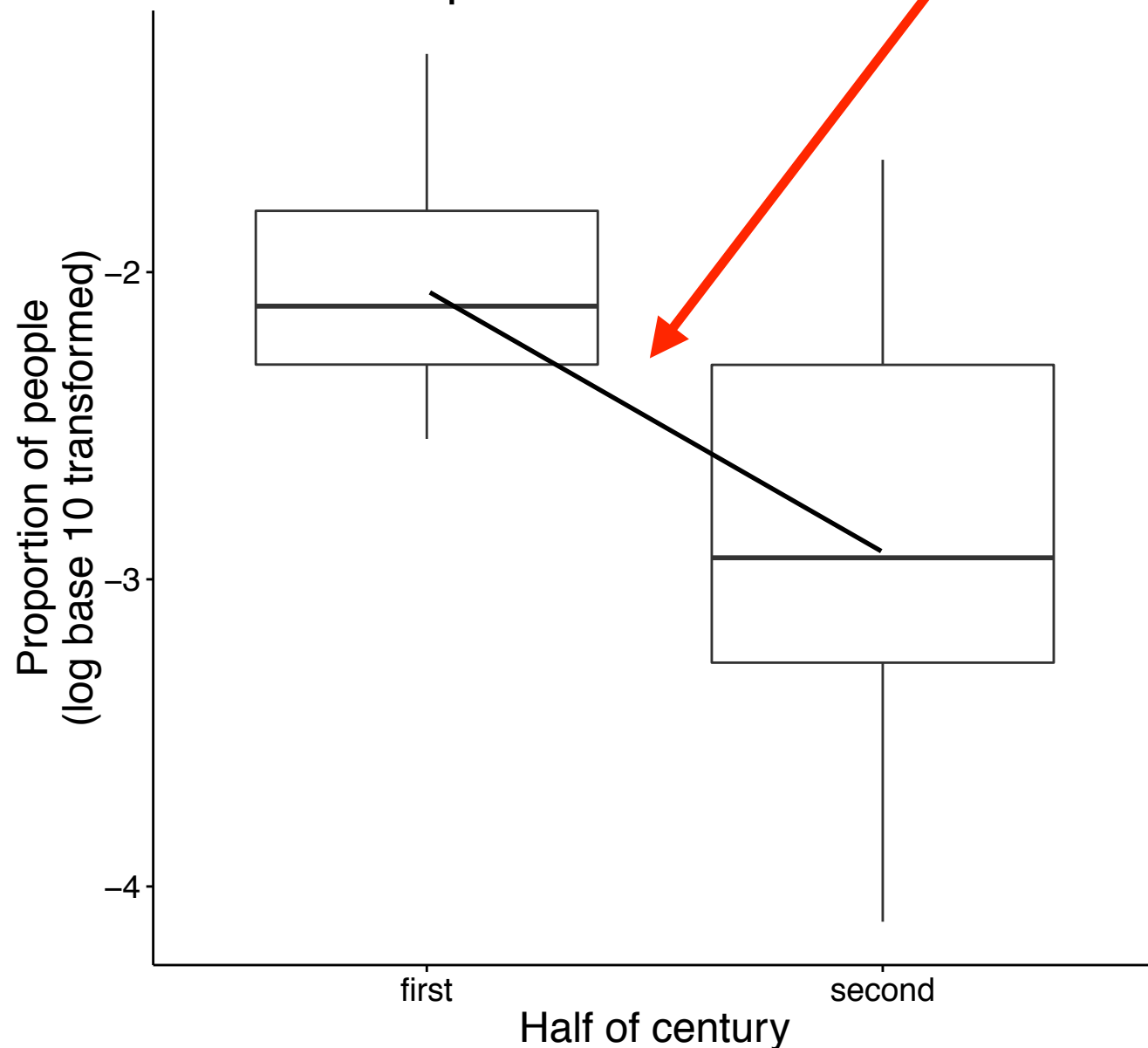


Group Means	Model
$x_1 0 x_{20}$	a
$x_1 1 x_{20}$	$a + b_1$
$x_1 0 x_{21}$	$a + b_2$
$x_1 1 x_{21}$	$a + b_1 + b_2 + b_3$

$$b_1 X_{1i}$$

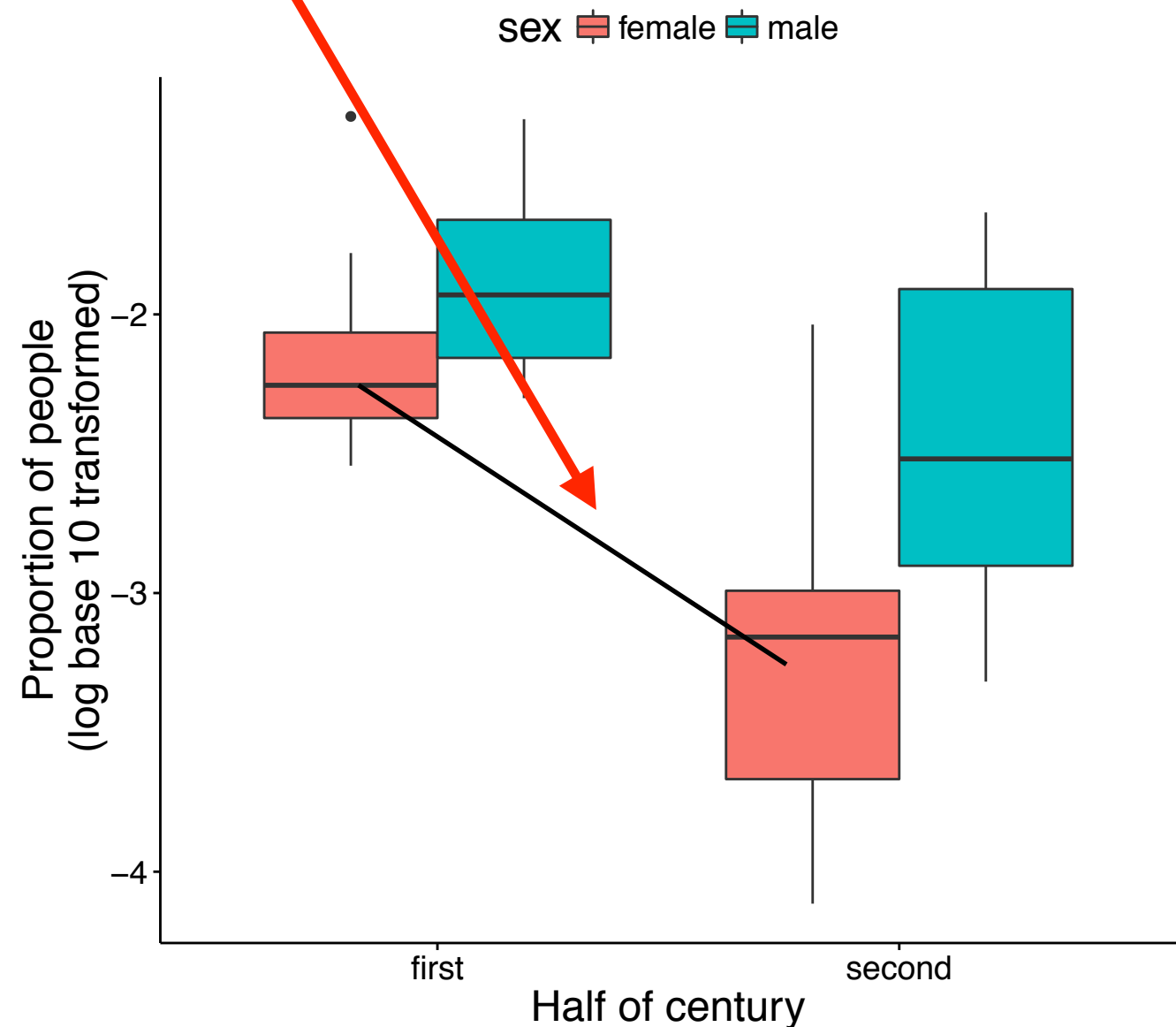
linear model without interaction

Proportion of People with
Popular Names for 1901



linear model with interaction

Proportion of People with
Popular Names for 1901



R Code (Part 2)

$$y_i = a + b_1X_{1i} + b_2X_{2i} + b_3X_{1i}X_{2i} + e_i$$

`lm(prop_log10_mean ~ century_half * sex)`

```
Call:
lm(formula = prop_log10_mean ~ century_half * sex, data = data_names)

Residuals:
    Min       1Q   Median       3Q      Max
-0.92578 -0.31259 -0.03546  0.26132  1.15233

Coefficients:
            Estimate Std. Error t value Pr(>|t|)
(Intercept)    -2.1569     0.1038  -20.875  < 2e-16 ***
century_halfsecond -1.022     0.1468   -6.962 1.04e-09 ***
sexmale         0.2967     0.1468    2.021  0.0468 *
century_halfsecond:sexmale 0.4318     0.2076    2.080  0.0409 *
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 0.4642 on 76 degrees of freedom
Multiple R-squared:  0.5395,    Adjusted R-squared:  0.5213
F-statistic: 29.67 on 3 and 76 DF,  p-value: 8.345e-13
```

`> head(resid(summary(popnames_interaction.lm)))`

1	2	3
-0.28065156	-0.43747476	0.06919675
4	5	6
0.06507396	0.07114409	0.61794560

$$y_i = a + b_1X_{1i} + b_2X_{2i} + b_3X_{1i}X_{2i} + e_i$$

```
lm(prop_log10_mean ~ century_half * sex)
```

```
Call:
lm(formula = prop_log10_mean ~ century_half * sex, data = data_names)

Residuals:
    Min       1Q   Median       3Q      Max
-0.92578 -0.31259 -0.03546  0.26132  1.15233

Coefficients:
            1            2            3
-0.28065156 -0.43747476  0.06919675
            4            5            6
 0.06507396  0.07114409  0.61794560

> head(resid(summary(popnames_interaction.lm)))

Estimate Std. Error t value Pr(>|t|)
(Intercept)      -2.1669    0.1038 -20.875  < 2e-16 ***
century_halfsecond -1.0220    0.1468  -6.962 1.04e-09 ***
sexmale           0.2967    0.1468   2.021  0.0468 *
century_halfsecond:sexmale 0.4318    0.2076   2.080  0.0409 *
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 0.4642 on 76 degrees of freedom
Multiple R-squared:  0.5395,    Adjusted R-squared:  0.5213
F-statistic: 29.67 on 3 and 76 DF,  p-value: 8.345e-13
```

$$y_i = a + b_1X_{1i} + b_2X_{2i} + b_3X_{1i}X_{2i} + e_i$$

`lm(prop_log10_mean ~
relevel(century_half, "second") * sex)`

Call:

```
lm(formula = prop_log10_mean ~ relevel(century_half, "second") *  
sex, data = data_names)
```

Residuals:

Min	1Q	Median	3Q	Max
-0.92578	-0.31259	-0.03546	0.26132	1.15233

Coefficients:

	Estimate	Std. Error	t value	Pr(> t)
(Intercept)	-2.7889	0.1038	-30.721	< 2e-16 ***
relevel(century_half, "second")first	1.9220	0.1468	6.962	1.04e-09 ***
sexmale	0.7735	0.1468	4.963	4.15e-06 ***
relevel(century_half, "second")first:sexmale	-0.4318	0.2076	-2.080	0.0409 *

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

Residual standard error: 0.4642 on 76 degrees of freedom

Multiple R-squared: 0.5395, Adjusted R-squared: 0.5213

F-statistic: 29.67 on 3 and 76 DF, p-value: 8.345e-13

```
> head(resid(popnames_interaction_second.lm))
```

1	2	3
-0.28065156	-0.43747476	0.06919675
4	5	6
0.06507396	0.07114409	0.61794560

multiple regression without interaction

	Estimate	Std. Error	t value	Pr(> t)	
(Intercept)	-2.27486	0.09182	-24.776	< 2e-16	***
century_halfsecond	-0.80612	0.10602	-7.603	5.88e-11	***
sexmale	0.51263	0.10602	4.835	6.66e-06	***

multiple regression with interaction

	Estimate	Std. Error	t value	Pr(> t)	
(Intercept)	-2.1669	0.1038	-20.875	< 2e-16	***
century_halfsecond	-1.0220	0.1468	-6.962	1.04e-09	***
sexmale	0.2967	0.1468	2.021	0.0468	*
century_halfsecond:sexmale	0.4318	0.2076	2.080	0.0409	*

multiple regression without interaction

	Estimate	Std. Error	t value	Pr(> t)	
(Intercept)	-2.27486	0.09182	-24.776	< 2e-16	***
century_halfsecond	-0.80612	0.10602	-7.603	5.88e-11	***
sexmale	0.51263	0.10602	4.835	6.66e-06	***

multiple regression with interaction

	Estimate	Std. Error	t value	Pr(> t)	
(Intercept)	-2.1669	0.1038	-20.875	< 2e-16	***
century_halfsecond	-1.0220	0.1468	-6.962	1.04e-09	***
sexmale	0.2967	0.1468	2.021	0.0468	*
century_halfsecond:sexmale	0.4318	0.2076	2.080	0.0409	*

multiple regression with interaction “century_half” releveled

	Estimate	Std. Error	t value	Pr(> t)	
(Intercept)	-3.1889	0.1038	-30.721	< 2e-16	***
relevel(century_half, "second")first	1.0220	0.1468	6.962	1.04e-09	***
sexmale	0.7285	0.1468	4.963	4.15e-06	***
relevel(century_half, "second")first:sexmale	-0.4318	0.2076	-2.080	0.0409	*

Lab



STAR TREK

Data set: Extinction Likelihood of Star Trek Alien Species

Series: Is a given species more or less likely to become extinct in “Star Trek: The Original Series” or “Star Trek: The Next Generation?”

Alignment: Is a given species more or less likely to become extinct if it is a friend or a foe of the Enterprise?

Series x Alignment: Is there an interaction between these variables?

y_i = likely to become extinct or not
 a = ? - will get from model
 b_1 = ? - will get from model
 x_1 = series
 b_2 = ? - will get from model
 x_2 = alignment

source: The Star Trek Project

dplyr, purrr

data =

dplyr, purrr

```
data = list.files(  
)
```

verb

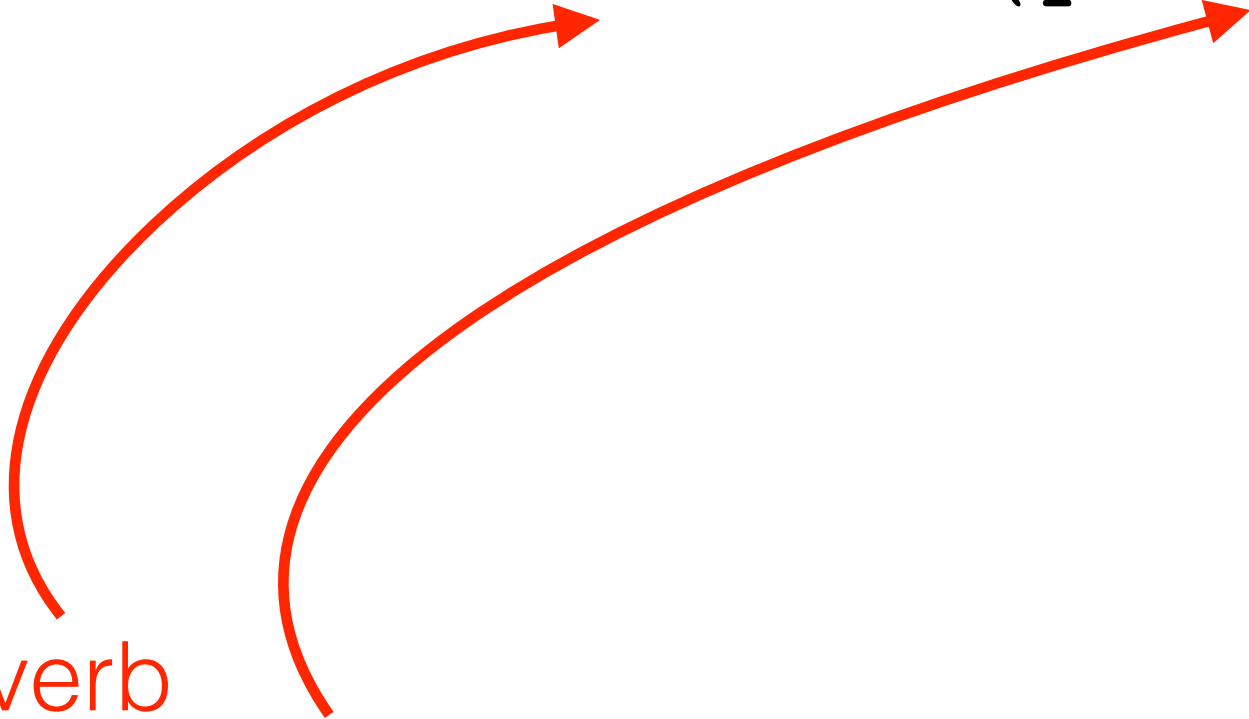


dplyr, purrr

```
data = list.files(path = "data")
```

verb

location
of files



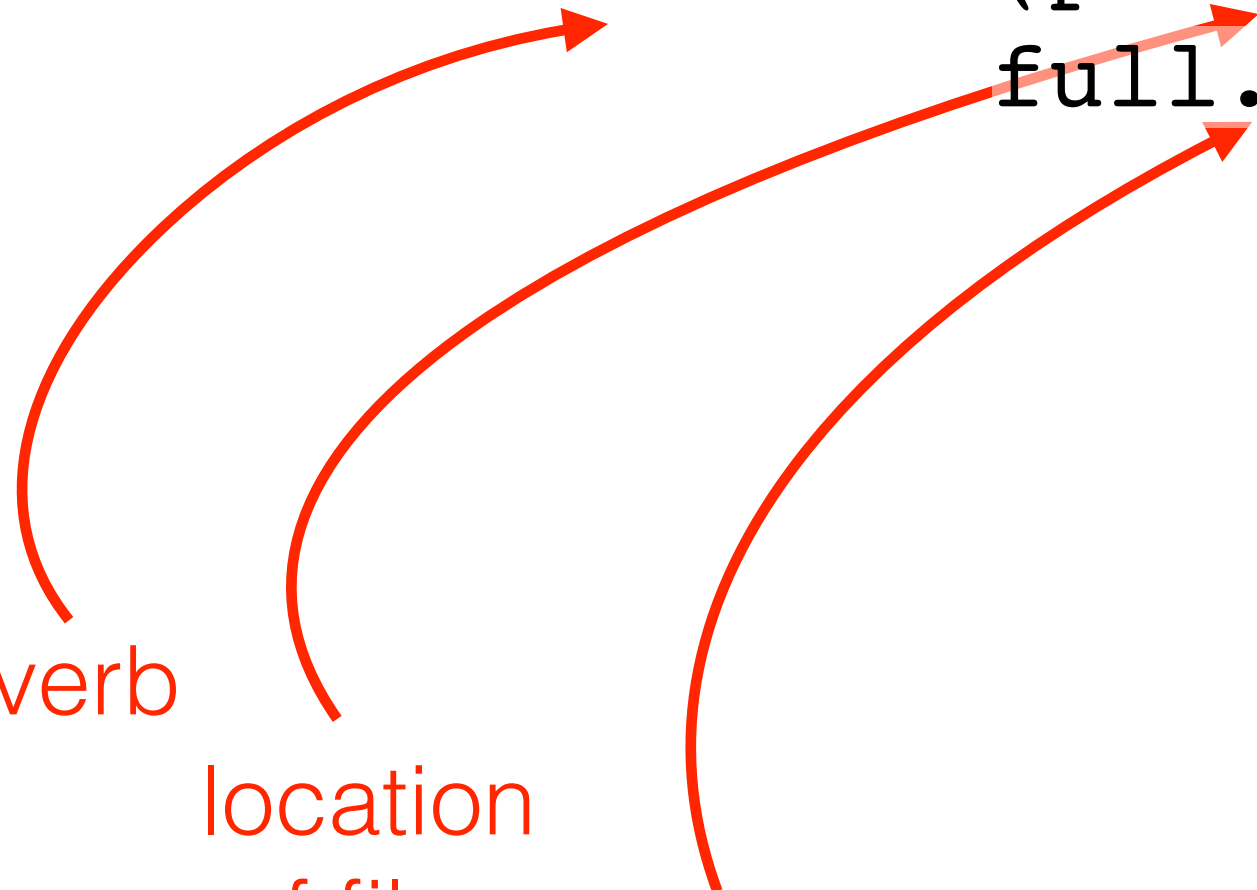
dplyr, purrr

```
data = list.files(path = "data",  
                  full.names = T)
```

verb

location
of files

how to
name files



dplyr, purrr

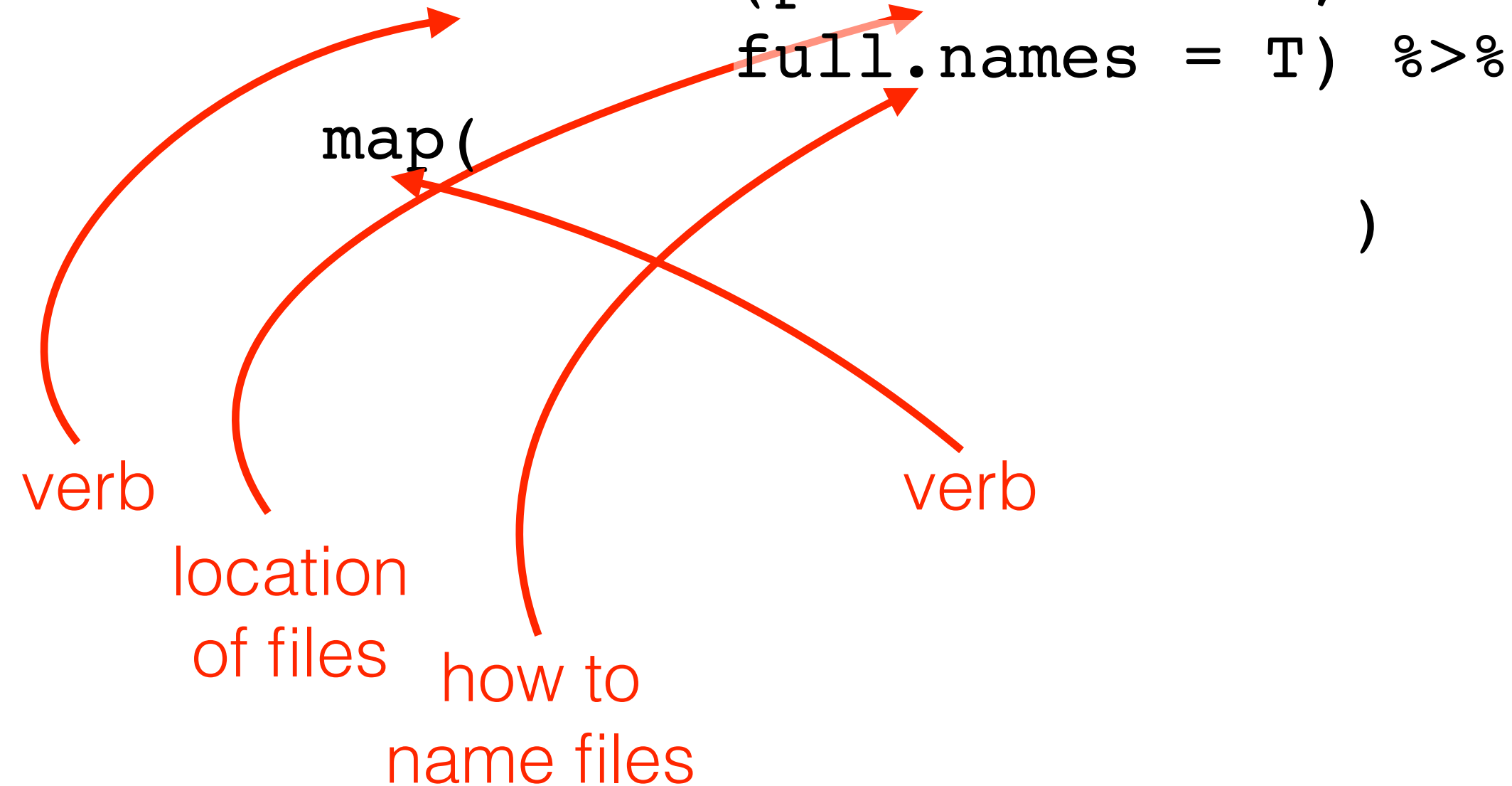
```
data = list.files(path = "data",  
                  full.names = T) %>%  
  map(  
    )
```

verb

location
of files

how to
name files

verb



dplyr, purrr

```
data = list.files(path = "data",  
                  full.names = T) %>%  
  map(read.table,  
      )
```

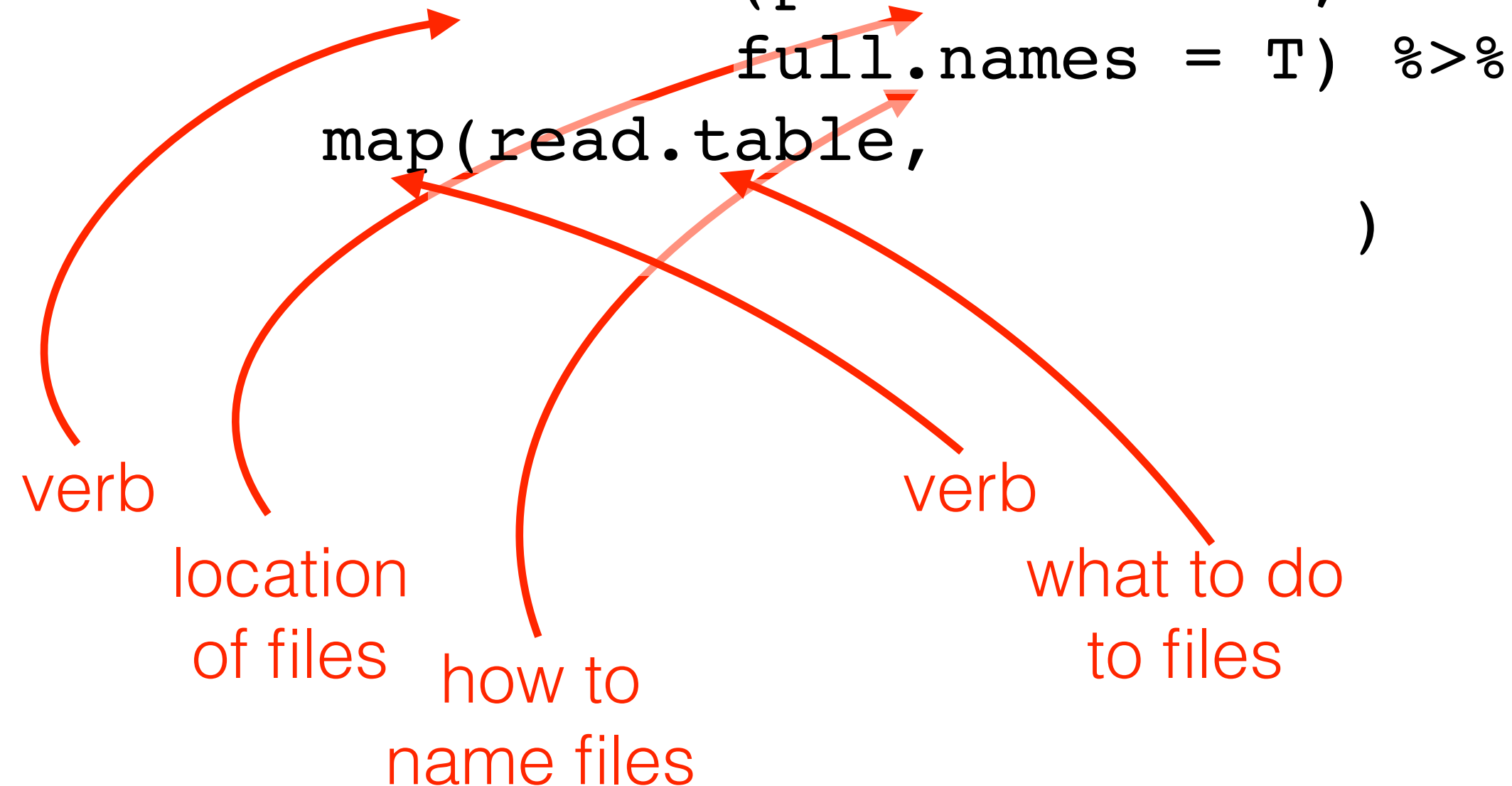
verb

location
of files

how to
name files

verb

what to do
to files



dplyr, purrr

```
data = list.files(path = "data",  
                  full.names = T) %>%  
  map(read.table, header = T, sep = "\t",  
       na.strings = c("", NA))
```

verb

location
of files

how to
name files

verb

what to do
to files

dplyr, purrr

```
data = list.files(path = "data",  
                  full.names = T) %>%  
  map(read.table, header = T, sep = "\t",  
       na.strings = c("", NA)) %>%  
  reduce( )
```

verb

location
of files

how to
name files

verb

what to do
to files

verb

dplyr, purrr

```
data = list.files(path = "data",  
                  full.names = T) %>%  
  map(read.table, header = T, sep = "\t",  
       na.strings = c("", NA)) %>%  
  reduce(rbind)
```

verb

location
of files

how to
name files

verb

what to do
to files

verb

what to do
to data frames

dplyr

```
data_clean = data %>%
```

dplyr

```
data_clean = data %>%  
  ...
```


dplyr

```
data_clean = data %>%
```

```
...
```

```
group_by(
```

```
)
```

verb



dplyr

```
data_clean = data %>%
```

```
...
```

```
group_by(series, alignment,  
alien)
```

verb

variables
to group by



dplyr

```
data_clean = data %>%
```

```
...
```

```
group_by(series, alignment,  
alien) %>%
```

```
arrange(  
)
```

verb

variables
to group by

verb

dplyr

```
data_clean = data %>%
```

```
...
```

```
group_by(series, alignment,  
alien) %>%
```

```
arrange(episode)
```

verb

variables
to group by

verb

variable to
order by
(ascending)

dplyr

```
data_clean = data %>%
```

```
...
```

```
group_by(series, alignment,
```

```
alien) %>%
```

```
arrange(episode) %>%
```

```
filter(
```

```
)
```

verb

variables
to group by

verb

variable to
order by
(ascending)

verb

dplyr

```
data_clean = data %>%
```

```
...
```

```
group_by(series, alignment,  
alien) %>%
```

```
arrange(episode) %>%
```

```
filter(row_number() == 1)
```

verb

variables
to group by

verb

variable to
order by
(ascending)

verb

rows
to keep

dplyr

```
data_clean = data %>%
```

```
...
```

```
group_by(series, alignment,  
alien) %>%
```

```
arrange(episode) %>%
```

```
filter(row_number() == 1) %>%
```

```
ungroup()
```

verb

variables
to group by

verb

variable to
order by
(ascending)

verb

rows
to keep

remove
grouping


ggplot2

`extinct.plot =`

ggplot2


```
extinct.plot = ggplot(data_figs,  
                      aes(x = series,  
                          y = perc_extinct,  
                          fill = alignment))
```

variable for
color fill



ggplot2

```
extinct.plot = ggplot(data_figs,  
                      aes(x = series,  
                          y = perc_extinct,  
                          fill = alignment)) +  
  geom_bar(stat = "identity",  
           position = "dodge")
```



variable for
color fill

make bars
side-by-side

ggplot2

```
extinct.plot = ggplot(data_figs,  
                      aes(x = series,  
                        y = perc_extinct,  
                        fill = alignment)) +  
  geom_bar(stat = "identity",  
          position = "dodge") +  
  ylim(0, 100) +  
  geom_hline(
```

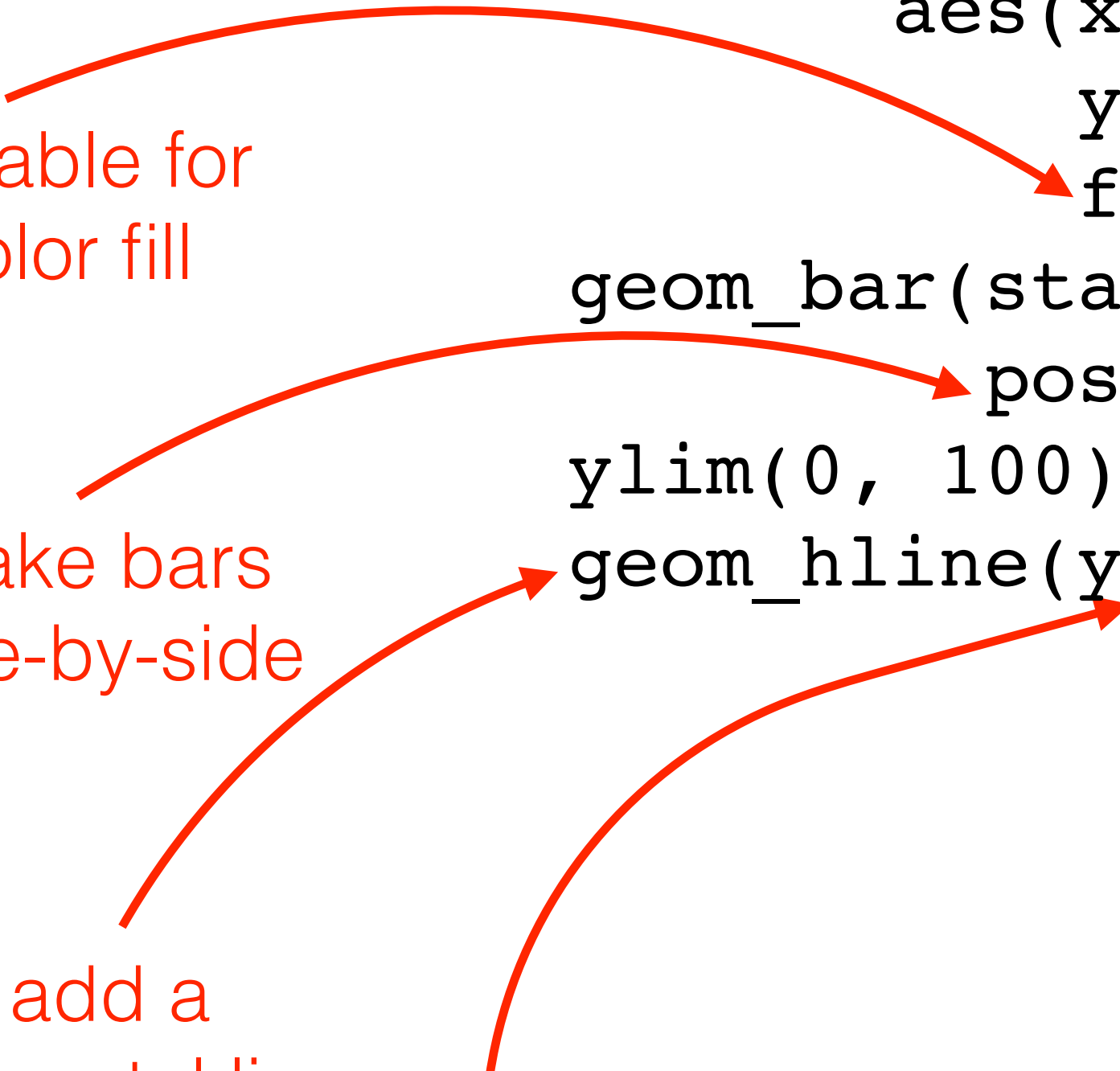
variable for
color fill

make bars
side-by-side

add a
horizontal line

ggplot2

```
extinct.plot = ggplot(data_figs,  
                      aes(x = series,  
                        y = perc_extinct,  
                        fill = alignment)) +  
  geom_bar(stat = "identity",  
          position = "dodge") +  
  ylim(0, 100) +  
  geom_hline(yintercept = 50)
```



variable for
color fill

make bars
side-by-side

add a
horizontal line

location
of line

ggplot2

```
extinct.plot = ggplot(data_figs,  
                      aes(x = series,  
                        y = perc_extinct,  
                        fill = alignment)) +  
  geom_bar(stat = "identity",  
          position = "dodge") +  
  ylim(0, 100) +  
  geom_hline(yintercept = 50) +  
  scale_fill_manual(  
    )
```

variable for
color fill

make bars
side-by-side

add a
horizontal line

location
of line

set colors
manually

ggplot2

```
extinct.plot = ggplot(data_figs,  
                      aes(x = series,  
                        y = perc_extinct,  
                        fill = alignment)) +  
  geom_bar(stat = "identity",  
          position = "dodge") +  
  ylim(0, 100) +  
  geom_hline(yintercept = 50) +  
  scale_fill_manual(  
    values = c("red", "yellow"))
```

variable for
color fill

make bars
side-by-side

add a
horizontal line

location
of line

set colors
manually

list colors