The SPINE of statistics: the linear model, parameters and estimation

Professor Andy Field

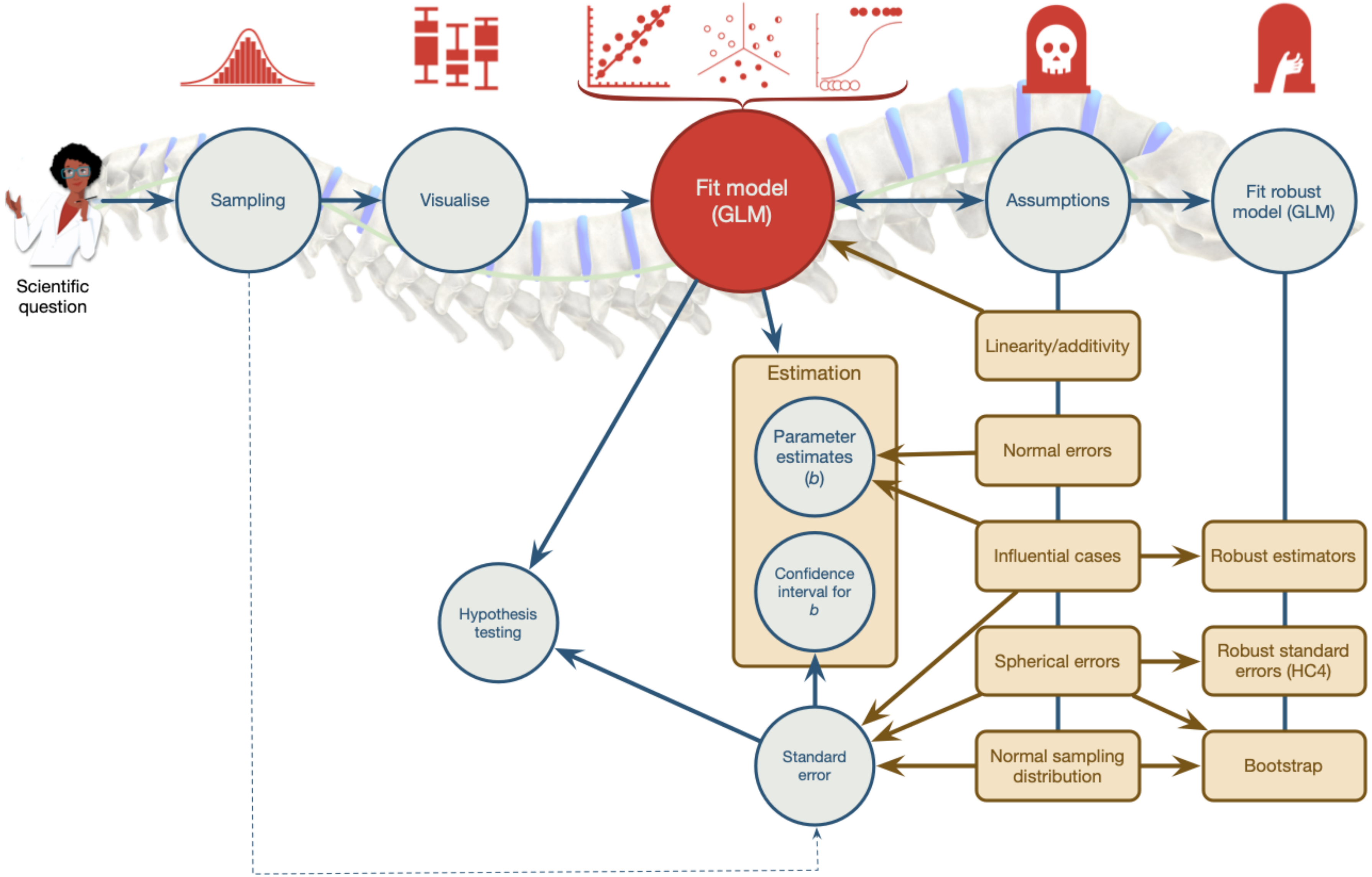
@profandyfield

www.youtube.com/user/ProfAndyField/

www.discoveringstatistics.com

www.milton-the-cat.rocks

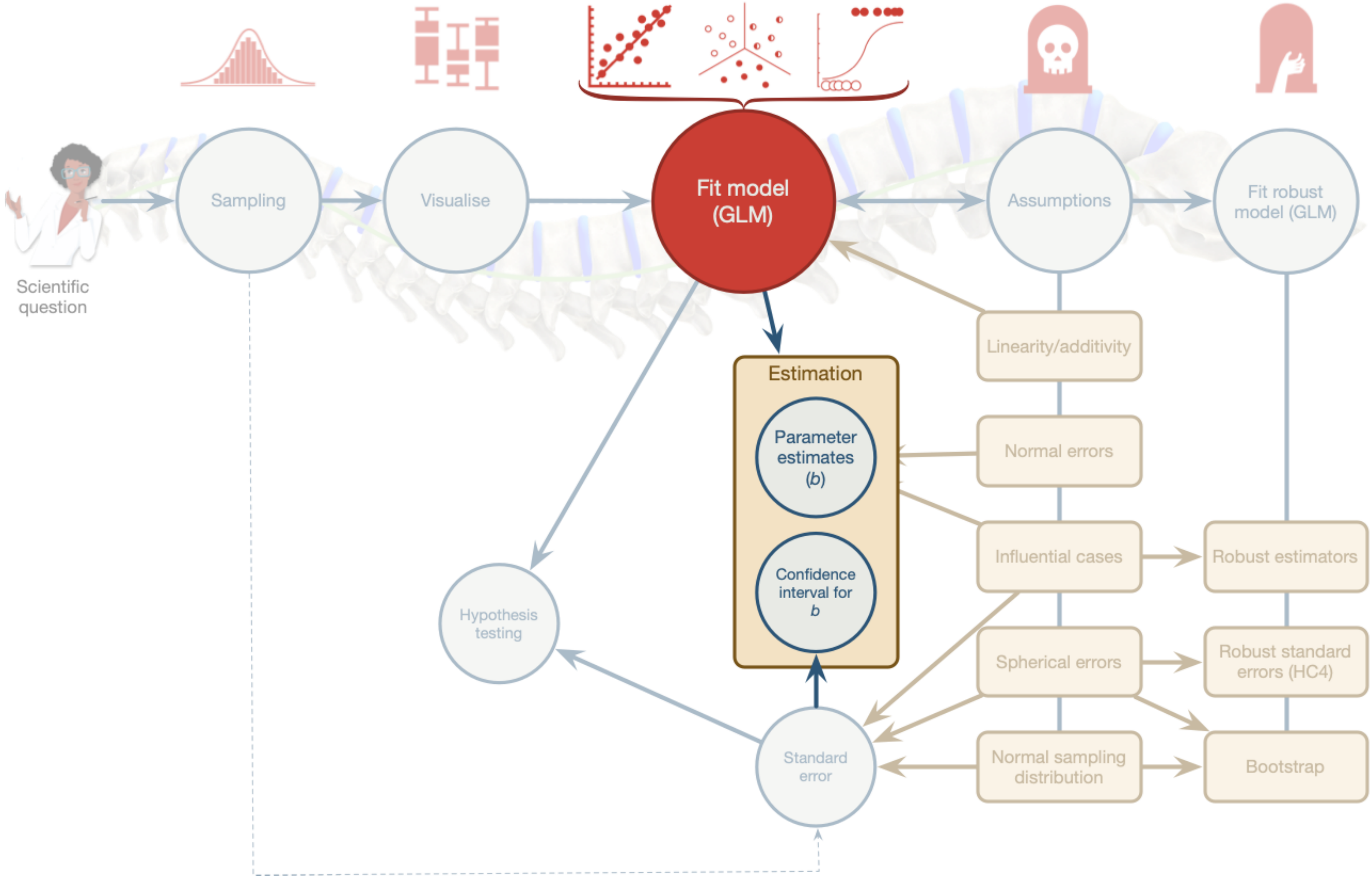
www.discovr.rocks


Scientific question
Sampling
Visualise
Fit model (GLM)
Assumptions
Fit robust model (GLM)
Estimation
Parameter estimates (b)
Confidence interval for b
Hypothesis testing
Standard error
Linearity/additivity
Normal errors
Influential cases
Spherical errors
Normal sampling distribution
Robust estimators
Robust standard errors (HC4)
Bootstrap

The SPINE of statistics

5 Key concepts

- Standard error
- Parameters
- Interval estimates
- Null hypothesis significance testing (NHST)
- Estimation


Scientific question
Sampling
Visualise
Fit model (GLM)
Assumptions
Fit robust model (GLM)
Estimation
Parameter estimates (b)
Confidence interval for b
Hypothesis testing
Standard error
Linearity/additivity
Normal errors
Influential cases
Spherical errors
Normal sampling distribution
Robust estimators
Robust standard errors (HC4)
Bootstrap

Learning outcomes

- Understand the commonalities in psychological statistical models
  - Most psychological statistical boil down to a very simple idea of predicting an outcome from one or more measured variables
- Understand the function and form of the linear model
  - Predicting an outcome variable from another variable (a predictor)
  - The mathematical model
  - Visualizing the model
  - Familiar models as variants of the linear model
- Understand what the model parameters (*b*s) represent
- Understand why we use sampling
- Understand (conceptually) least squares estimation

Why do (some) students hate statistics?

$$\hat{Y}_i = \hat{b}_0 + \hat{b}_1\mathrm{X}_i + \mathrm{e}_i$$

$$r = \frac{\sum_{i=1}^{n}(x_i - \bar{X})(y_i - \bar{Y})}{(N-1)s_x s_y}$$

$$t = \frac{\bar{X}_1 - \bar{X}_2}{\sqrt{\frac{s_1^2}{n_1} + \frac{s_2^2}{n_2}}}$$

$$\chi^2 = \sum \frac{(\mathrm{observed}_{ij} - \mathrm{model}_{ij})^2}{\mathrm{model}_{ij}}$$

Can we make statistics (a bit) less like this?

The only equation you will ever need

The General Linear Model (GLM)

$$\text{outcome}_i = (\text{model}_i) + \text{error}_i$$

$$\hat{\text{outcome}}_i = \hat{b}_0 + \hat{b}_1\text{predictor}_i + \cdots + \text{error}_i$$

A zombie quiz

> A researcher counted how many humans and zombies choose brain chips or potato chips to accompany their dinner at the university canteen

- How do I analyze these data?

Organism	Brain chips	Potato chips
Human	28	42
Zombie	61	57

A chi-square test

```
chisq.test(zom_fct_tib$organism, zom_fct_tib$chip, correct = FALSE)
```

```
##
##      Pearson's Chi-squared test
##
## data:  zom_fct_tib$organism and zom_fct_tib$chip
## X-squared = 2.4105, df = 1, p-value = 0.1205
```

A Spearman correlation?

```r
zom_tib %>%
  correlation::correlation(., method = "spearman")
```

Parameter1	Parameter2	rho	CI_low	CI_high	S	p	Method	n_Obs
organism	chip	-0.113	-0.252	0.03	1232810	0.122	Spearman	188

A Kendall's τ correlation?

```
zom_tib %>%
  correlation::correlation(., method = "kendall")
```

Parameter1	Parameter2	CI_low	CI_high	tau	z	p	Method	n_Obs
organism	chip	-0.252	0.03	-0.113	-1.548	0.122	Kendall	188

A Pearson correlation?

```
zom_tib %>%
  correlation::correlation()
```

Parameter1	Parameter2	r	CI_low	CI_high	t	df	p	Method	n_Obs
organism	chip	-0.113	-0.252	0.03	-1.554	186	0.122	Pearson	188

A *t*-test?

```
brains <- zom_tib %>% dplyr::filter(chip == 0)
potatoes <- zom_tib %>% dplyr::filter(chip == 1)
```

```
t.test(brains$organism, potatoes$organism)
```

estimate	estimate1	estimate2	statistic	p.value	parameter	conf.low	conf.high	method	alternative
0.11	0.685	0.576	1.559	0.121	185.619	-0.029	0.248	Welch Two Sample t-test	two.sided

One-way ANOVA?

```
zom_tib %>%
  aov(organism ~ factor(chip), data = .)
```

term	df	sumsq	meansq	statistic	p.value
factor(chip)	1	0.563	0.563	2.416	0.122
Residuals	186	43.373	0.233	NA	NA

Linear model (Regression)?

```
zom_tib %>%
  lm(organism ~ factor(chip), data = .)
```

term	estimate	std.error	statistic	p.value
(Intercept)	0.685	0.051	13.390	0.000
factor(chip)1	-0.110	0.071	-1.554	0.122

Loglinear model?

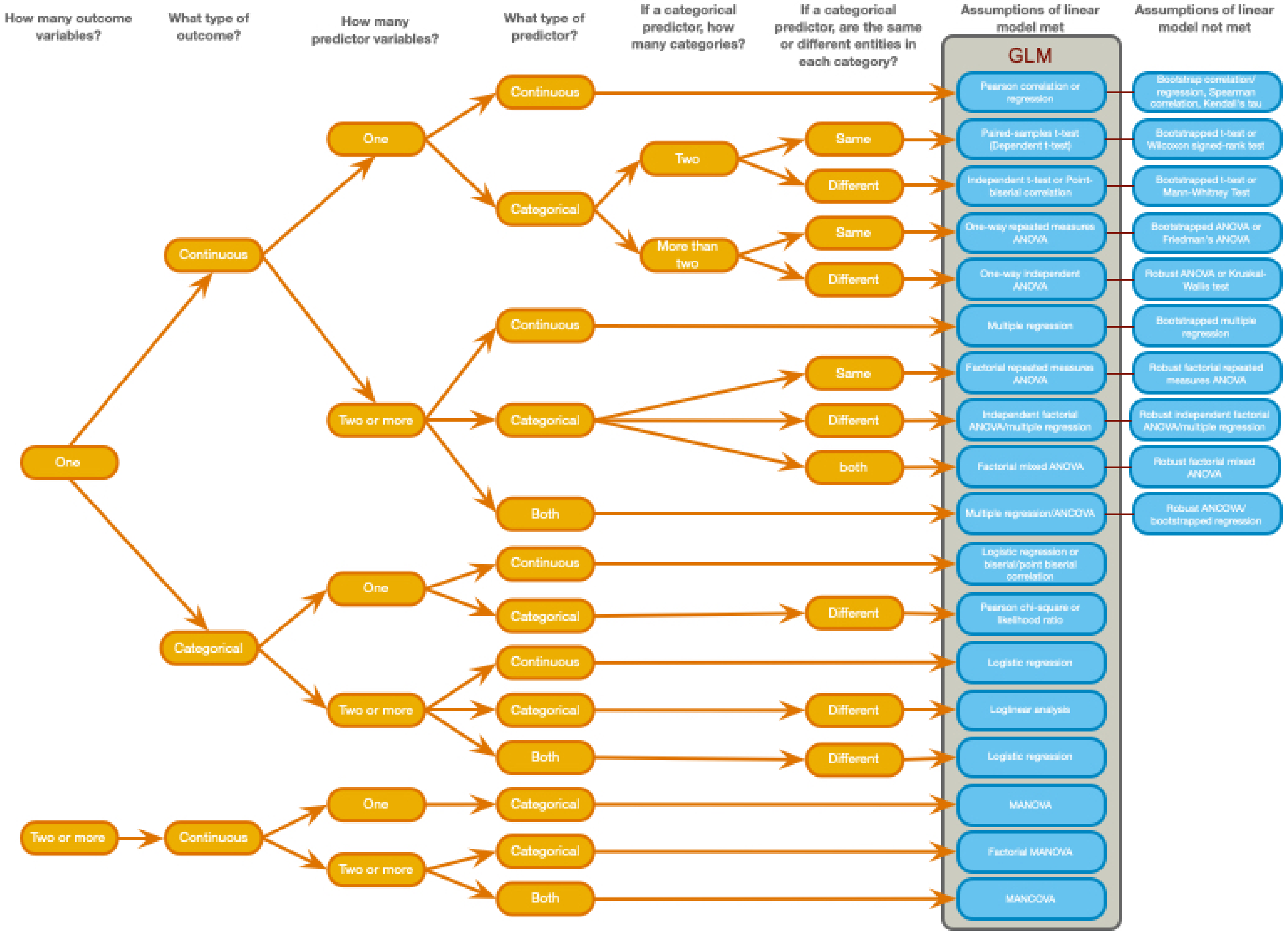
```
xtabs(~ organism + chip, data = zom_fct_tib) %>%
  MASS::loglm(~ organism + chip, data = .)
```

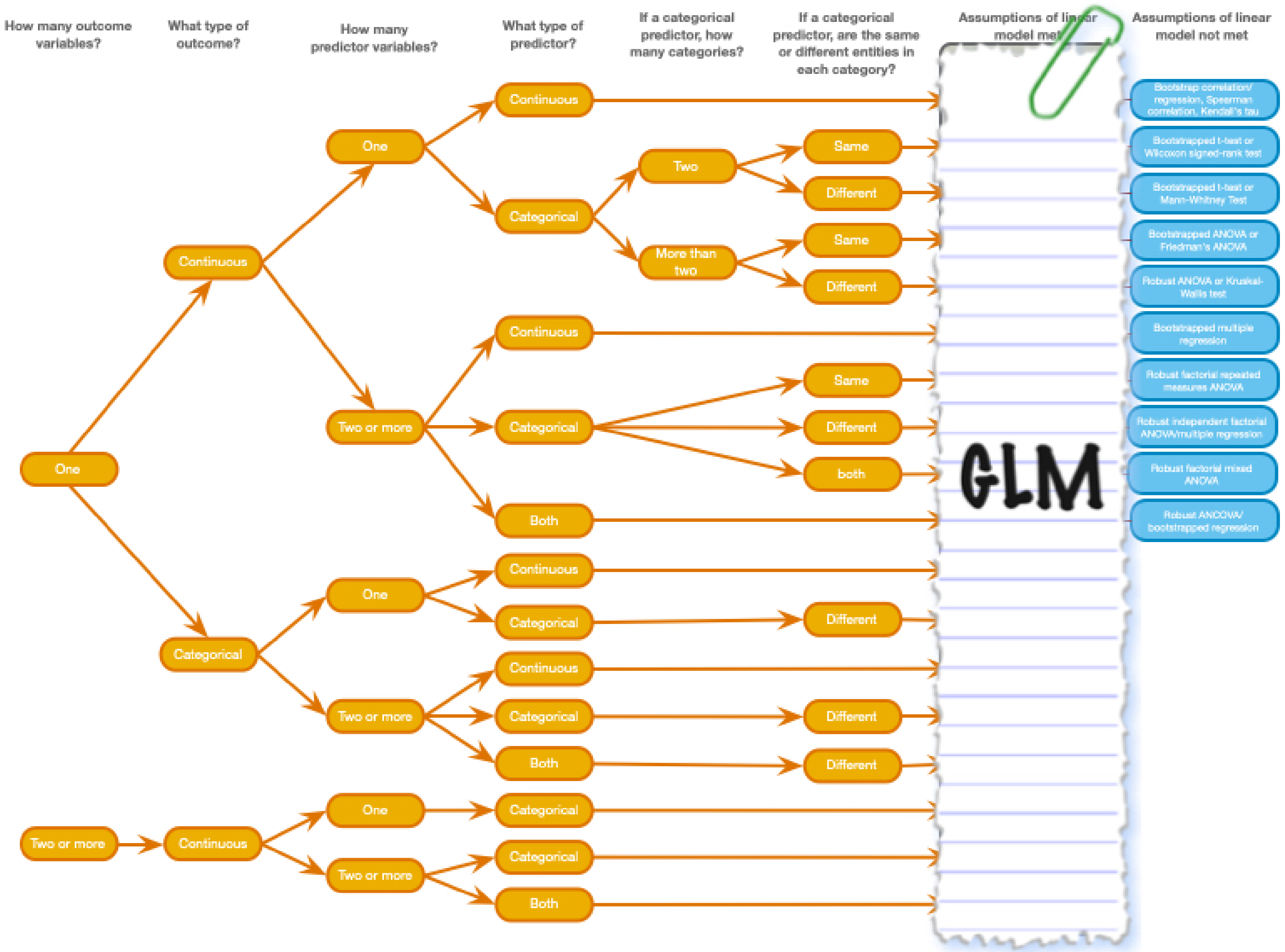
X^2	df	P(> X^2)
2.422	1	0.120
2.410	1	0.121

Multilevel model?

```
lmerTest::lmer(chip ~ organism + (1|canteen), data = zom_tib)
```

Parameter	Coefficient	SE	CI_low	CI_high	t	df_error	p
(Intercept)	0.600	0.060	0.483	0.717	10.065	184	0.00
organism	-0.117	0.075	-0.264	0.031	-1.554	184	0.12


How many outcome variables?
What type of outcome?
How many predictor variables?
What type of predictor?
If a categorical predictor, how many categories?
If a categorical predictor, are the same or different entities in each category?
Assumptions of linear model met
Assumptions of linear model not met
GLM
One
Continuous
One
Continuous
Categorical
Two
Same
Different
More than two
Same
Different
Two or more
Continuous
Categorical
Same
Different
both
Both
Categorical
One
Continuous
Categorical
Different
Two or more
Continuous
Categorical
Different
Both
Different
Two or more
Continuous
One
Categorical
Two or more
Categorical
Both
Pearson correlation or regression
Bootstrap correlation/ regression, Spearman correlation, Kendall's tau
Paired-samples t-test (Dependent t-test)
Bootstrapped t-test or Wilcoxon signed-rank test
Independent t-test or Point-biserial correlation
Bootstrapped t-test or Mann-Whitney Test
One-way repeated measures ANOVA
Bootstrapped ANOVA or Friedman's ANOVA
One-way independent ANOVA
Robust ANOVA or Kruskal-Wallis test
Multiple regression
Bootstrapped multiple regression
Factorial repeated measures ANOVA
Robust factorial repeated measures ANOVA
Independent factorial ANOVA/multiple regression
Robust independent factorial ANOVA/multiple regression
Factorial mixed ANOVA
Robust factorial mixed ANOVA
Multiple regression/ANCOVA
Robust ANCOVA/ bootstrapped regression
Logistic regression or biserial/point biserial correlation
Pearson chi-square or likelihood ratio
Logistic regression
Loglinear analysis
Logistic regression
MANOVA
Factorial MANOVA
MANCOVA



How many outcome variables?
What type of outcome?
How many predictor variables?
What type of predictor?
If a categorical predictor, how many categories?
If a categorical predictor, are the same or different entities in each category?
Assumptions of lin ar model met
Assumptions of linear model not met
One
Continuous
One
Continuous
Categorical
Two
Same
Different
More than two
Same
Different
Two or more
Continuous
Categorical
Same
Different
both
Both
Categorical
One
Continuous
Categorical
Different
Two or more
Continuous
Categorical
Different
Both
Different
Two or more
Continuous
One
Categorical
Two or more
Categorical
Both
GLM
Bootstrap correlation/ regression, Spearman correlation, Kendall's tau
Bootstrapped t-test or Wilcoxon signed-rank test
Bootstrapped t-test or Mann-Whitney Test
Bootstrapped ANOVA or Friedman's ANOVA
Robust ANOVA or Kruskal-Wallis test
Bootstrapped multiple regression
Robust factorial repeated measures ANOVA
Robust independent factorial ANOVA/multiple regression
Robust factorial mixed ANOVA
Robust ANCOVA/ bootstrapped regression

The only equation you will ever need

The General Linear Model (GLM)

$$\text{outcome}_i = (\text{model}_i) + \text{error}_i$$

$$\widehat{\text{outcome}}_i = \hat{b}_0 + \hat{b}_1\text{predictor}_i + \cdots + \hat{b}_n\text{predictor}_i + \text{error}_i$$

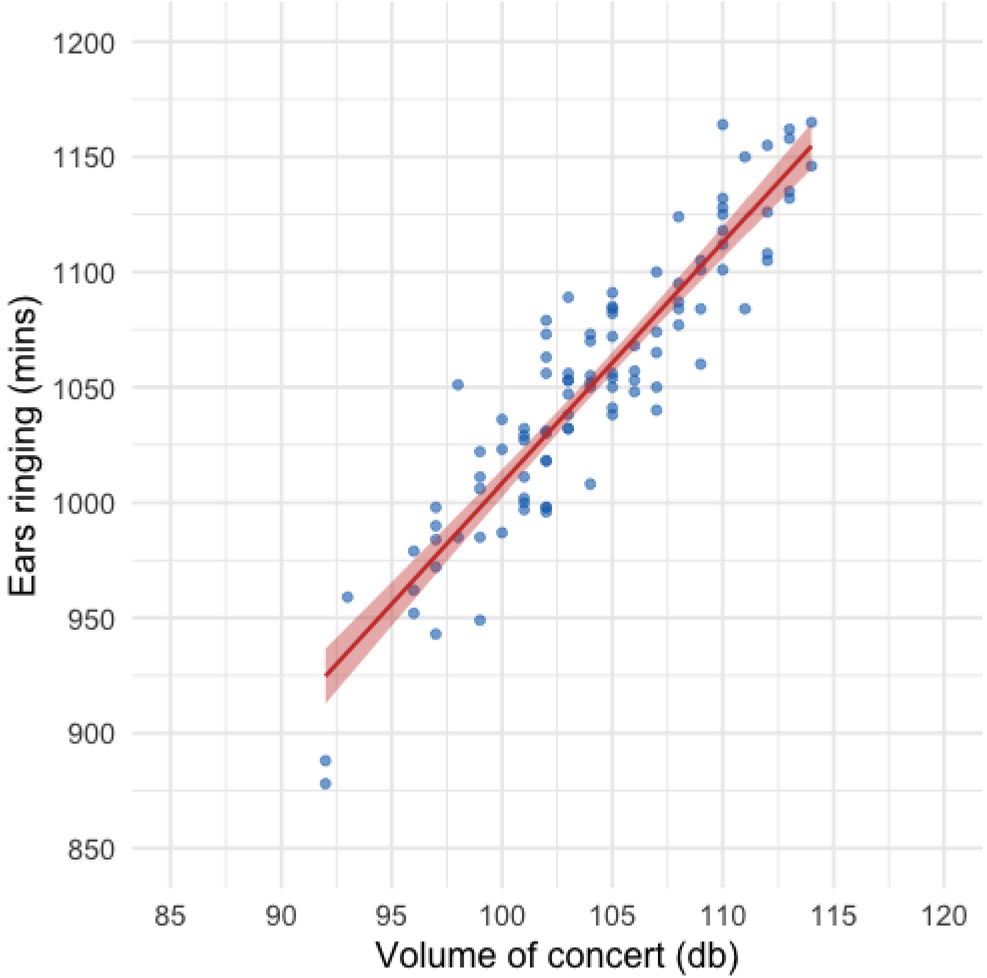
$\hat{b}_n$

- Estimate of parameter for a predictor
  - Direction/strength of relationship/effect
  - Difference in means

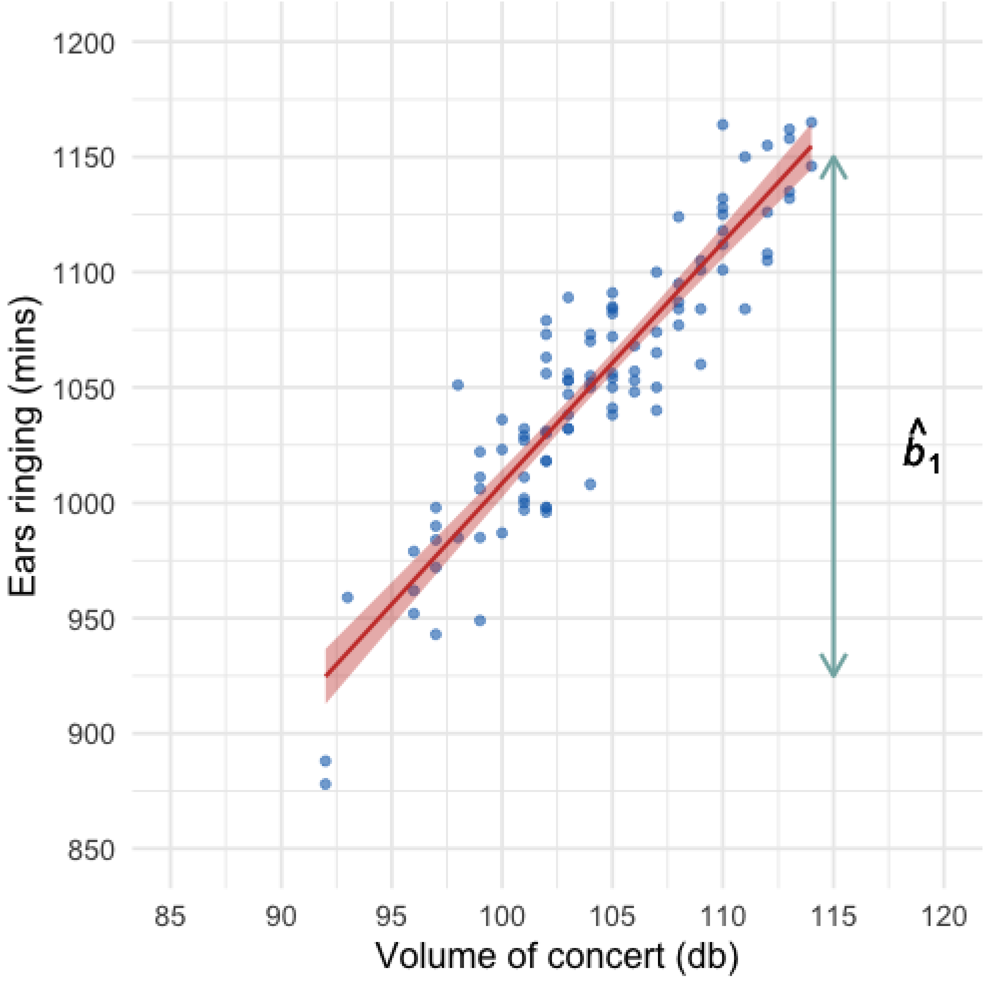
$\hat{b}_0$

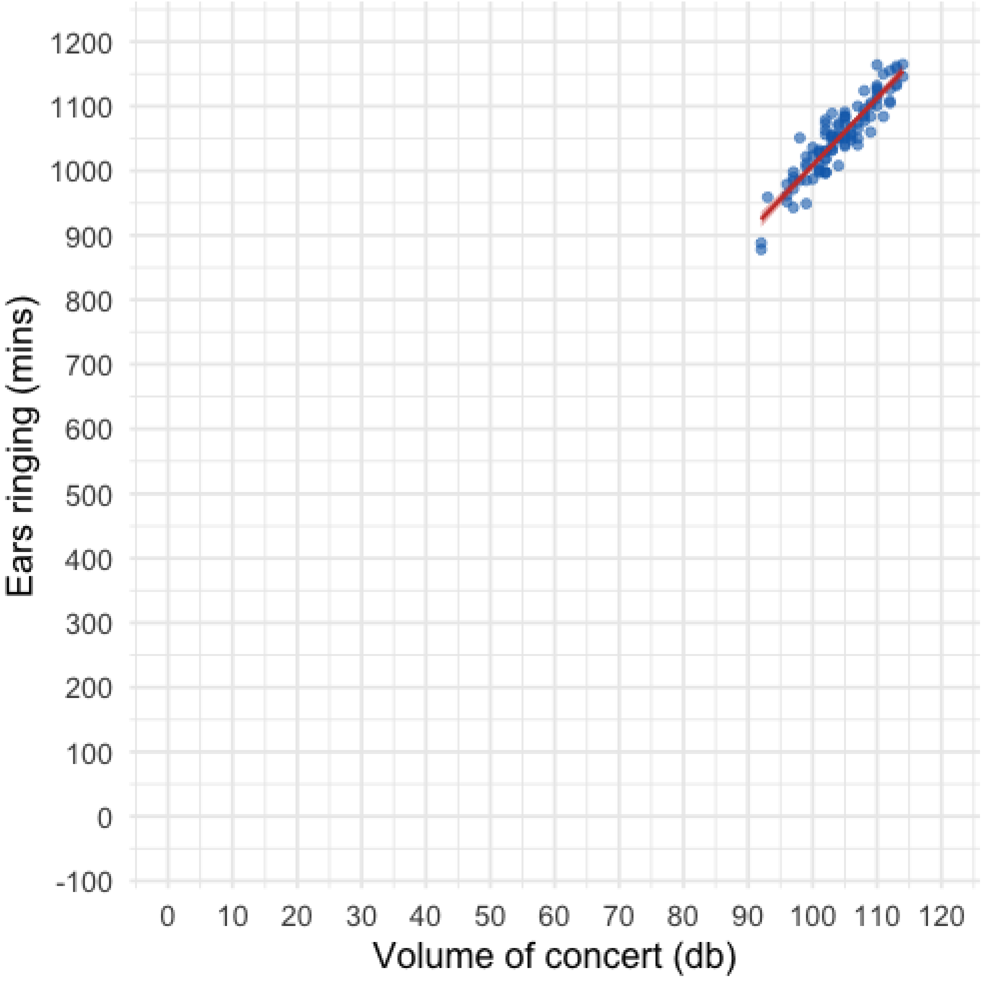
- Estimate of the value of the outcome when predictor(s) = 0 (intercept)

$$\hat{\text{ringing}}_i = \hat{b}_0 + \hat{b}_1 \text{volume}_i + e_i$$

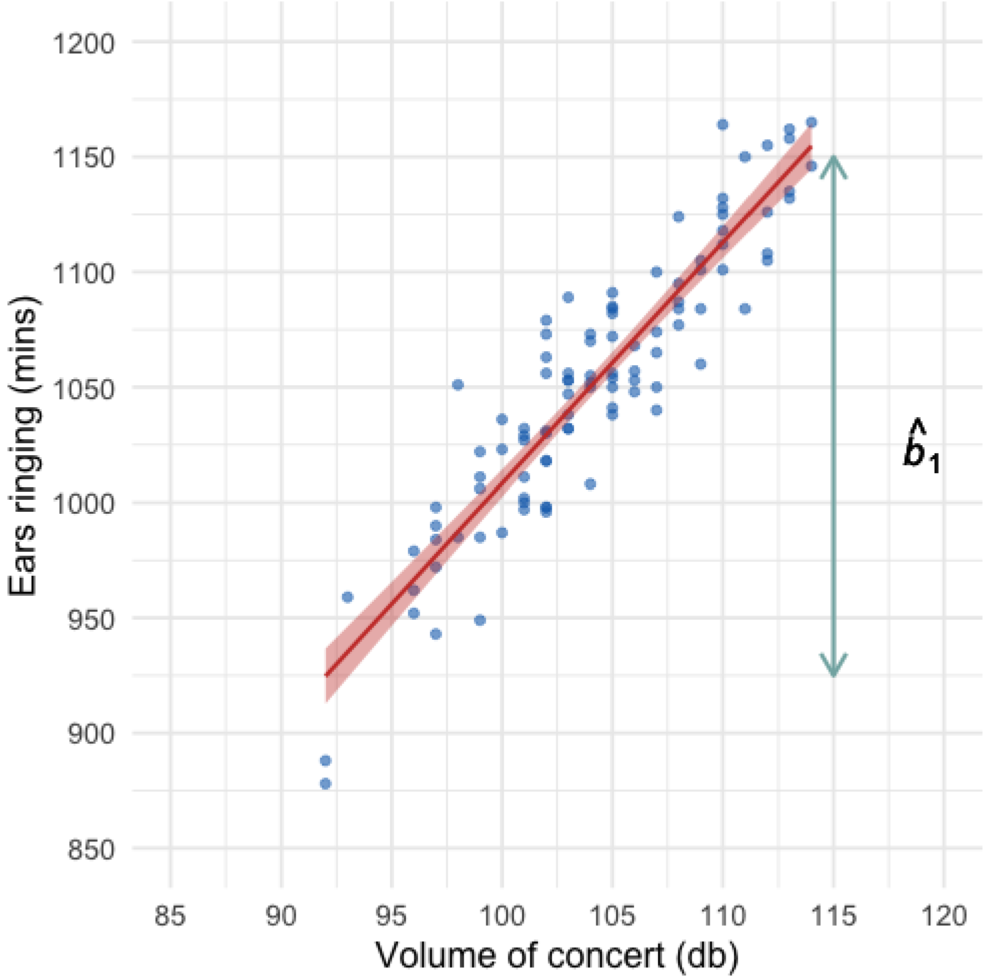


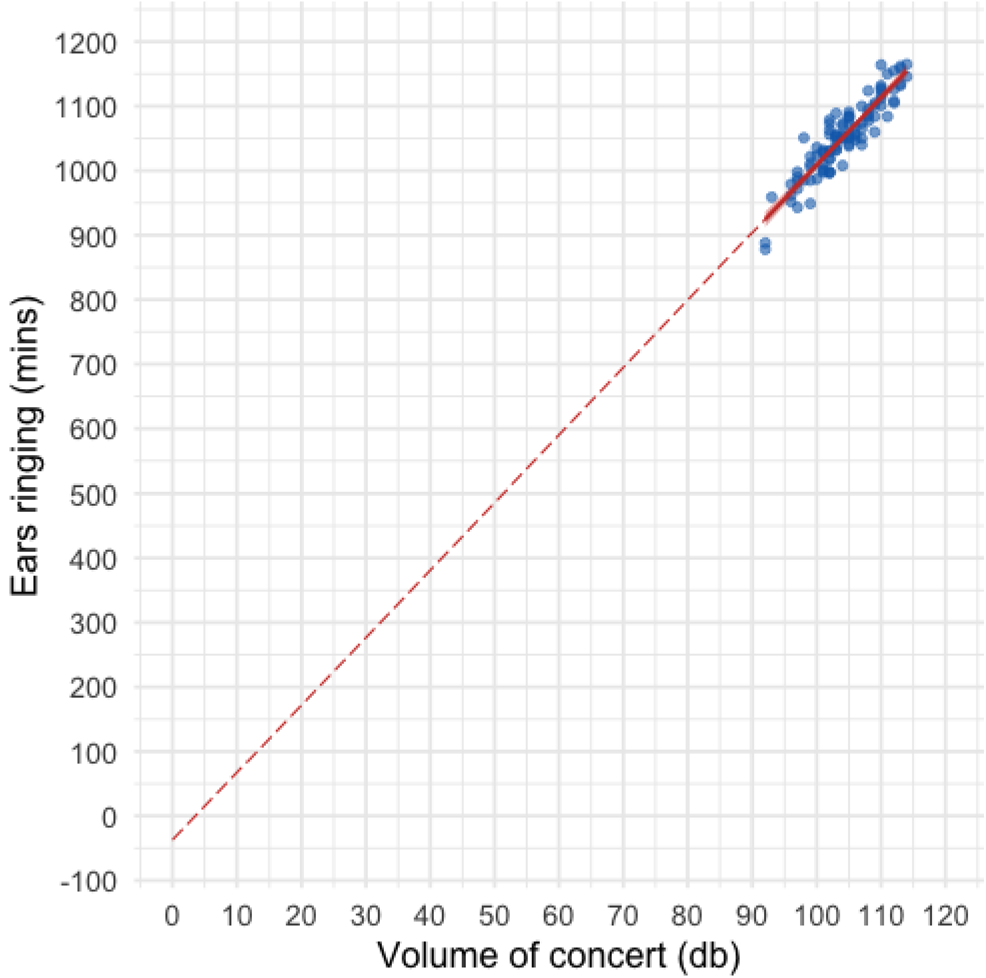
$$\hat{\text{ringing}}_i = \hat{b}_0 + \hat{b}_1 \text{volume}_i + e_i$$



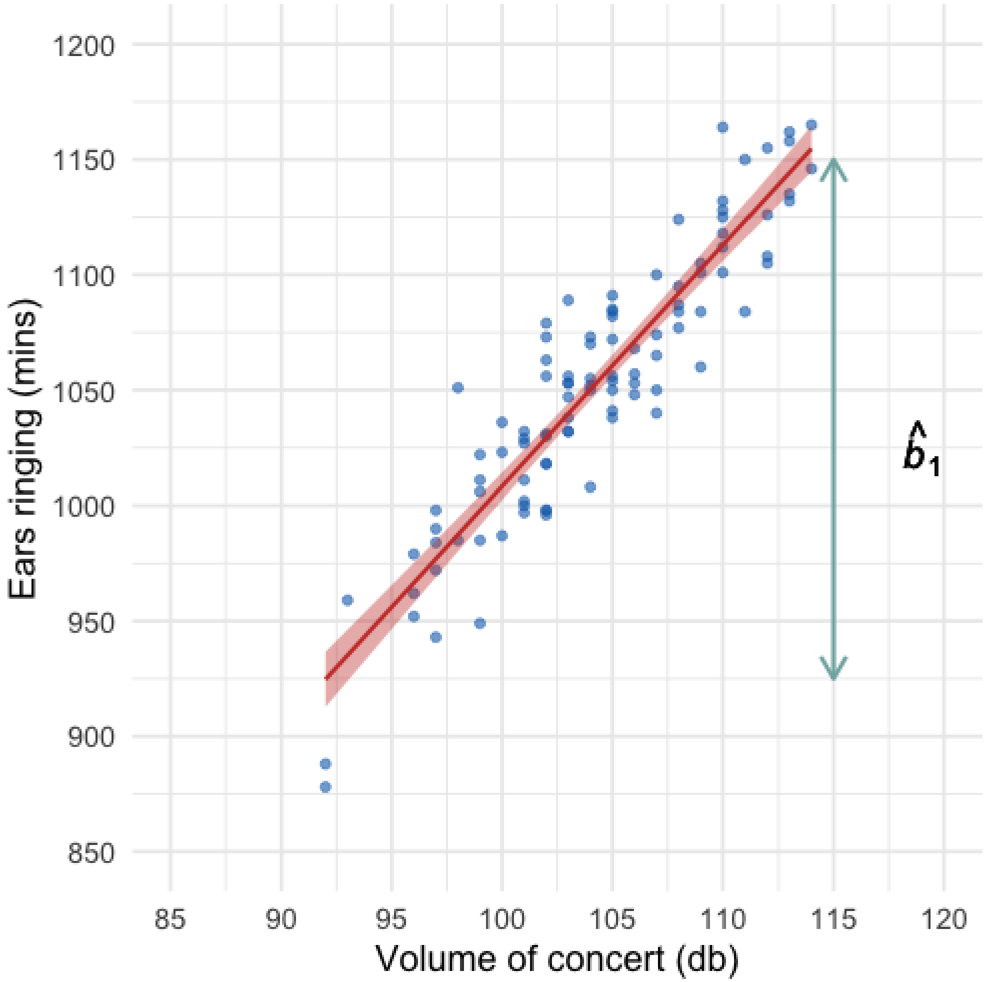


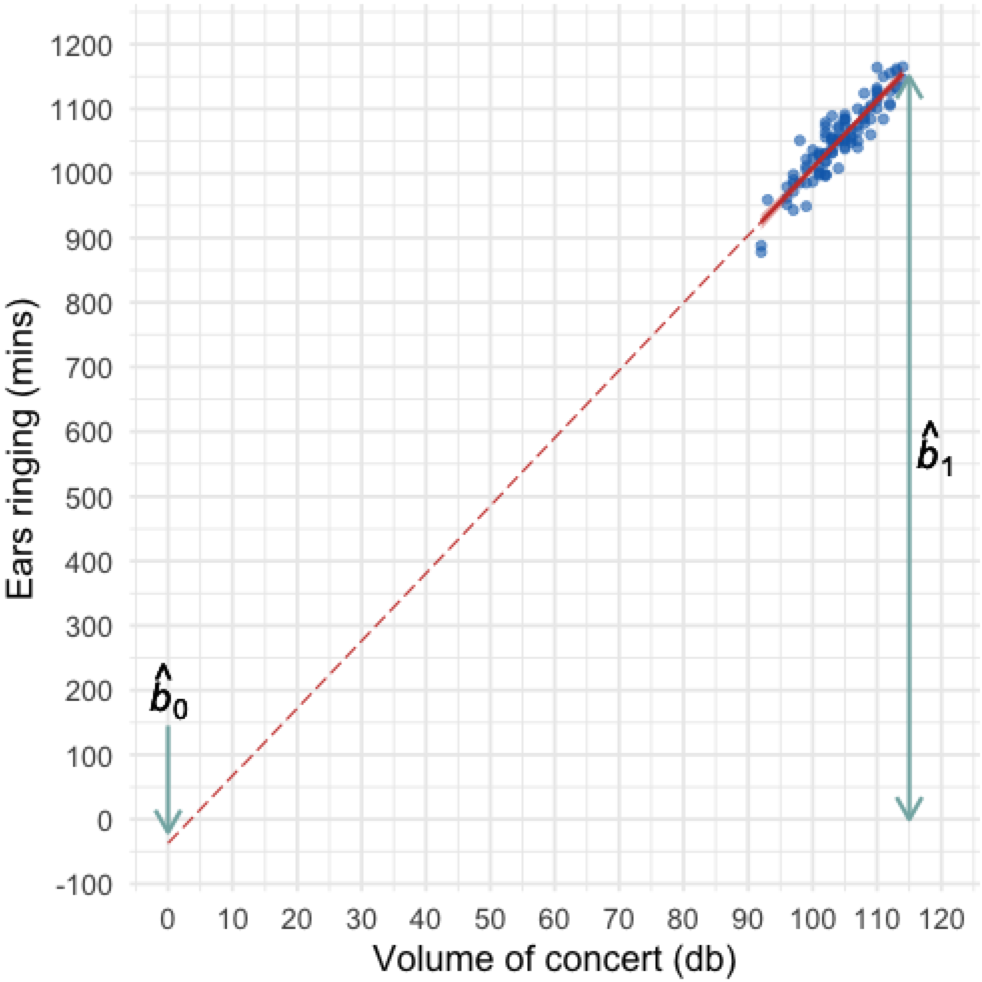
$$\hat{\text{ringing}}_i = \hat{b}_0 + \hat{b}_1\text{volume}_i + e_i$$



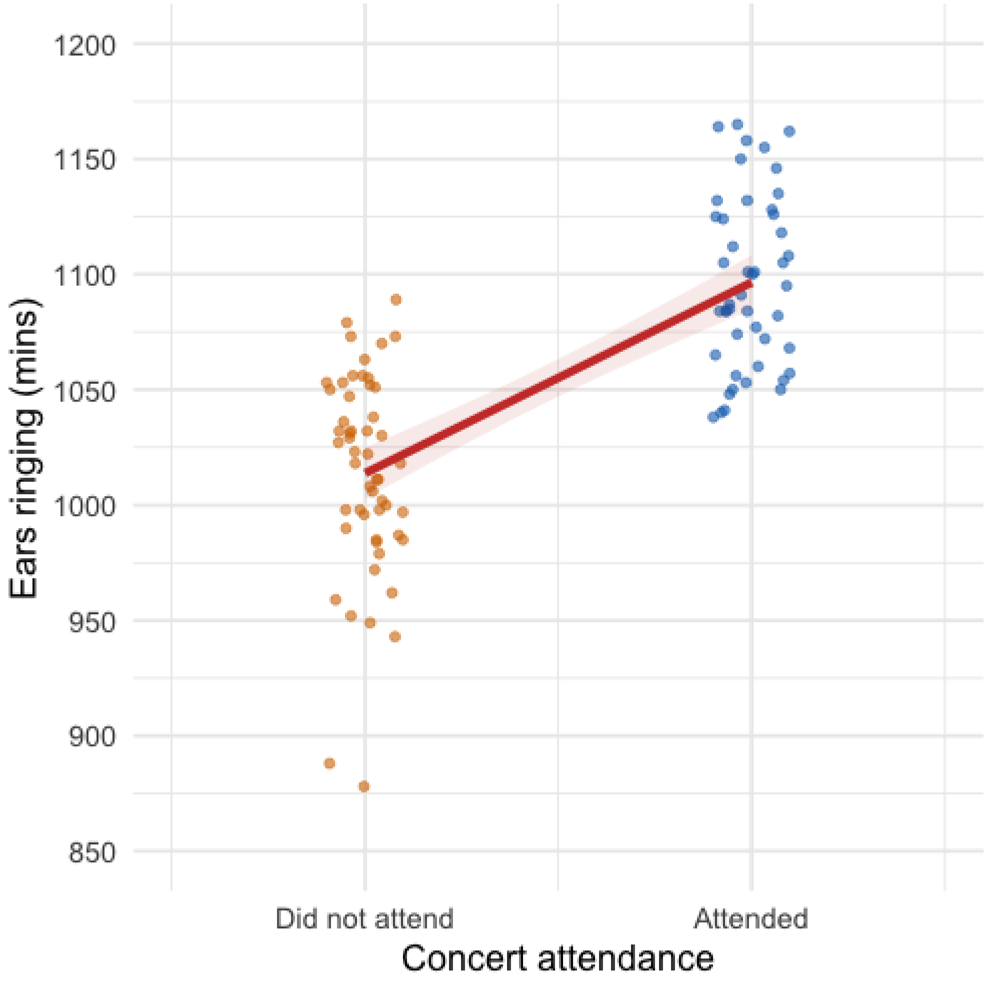


$$\hat{\text{ringing}}_i = -37.12 + 10.45\text{volume}_i + e_i$$

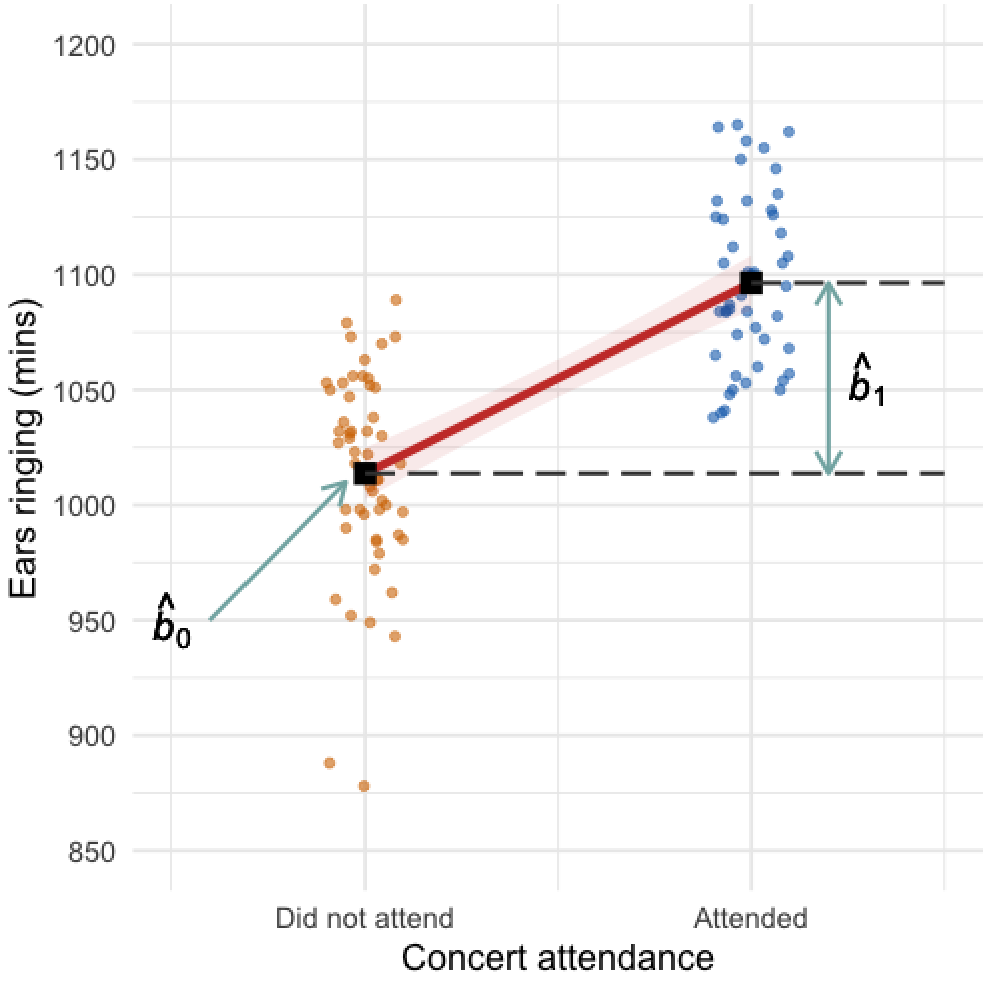




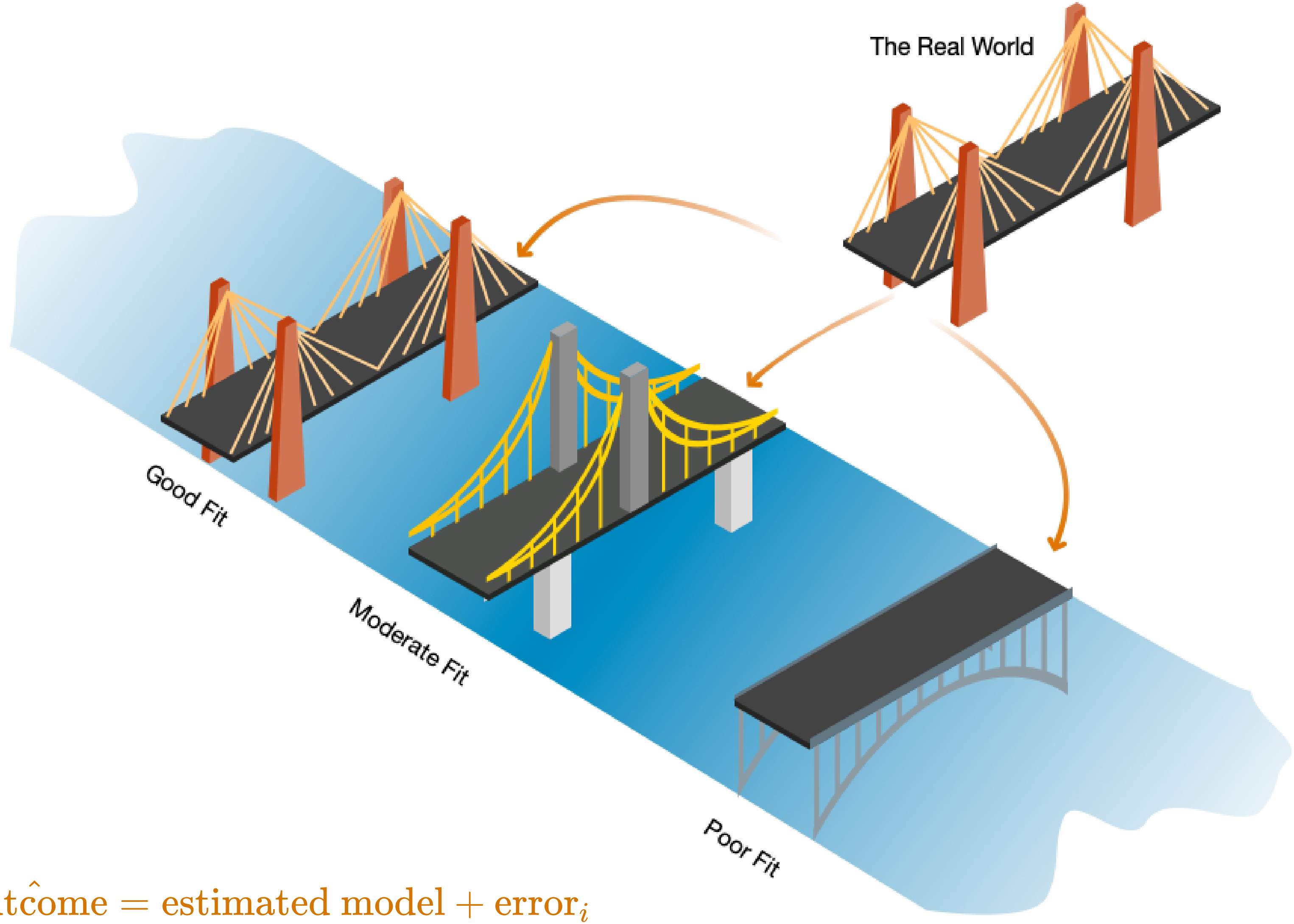
$$\hat{\text{ringing}}_i = \hat{b}_0 + \hat{b}_1\text{attendance}_i + e_i$$



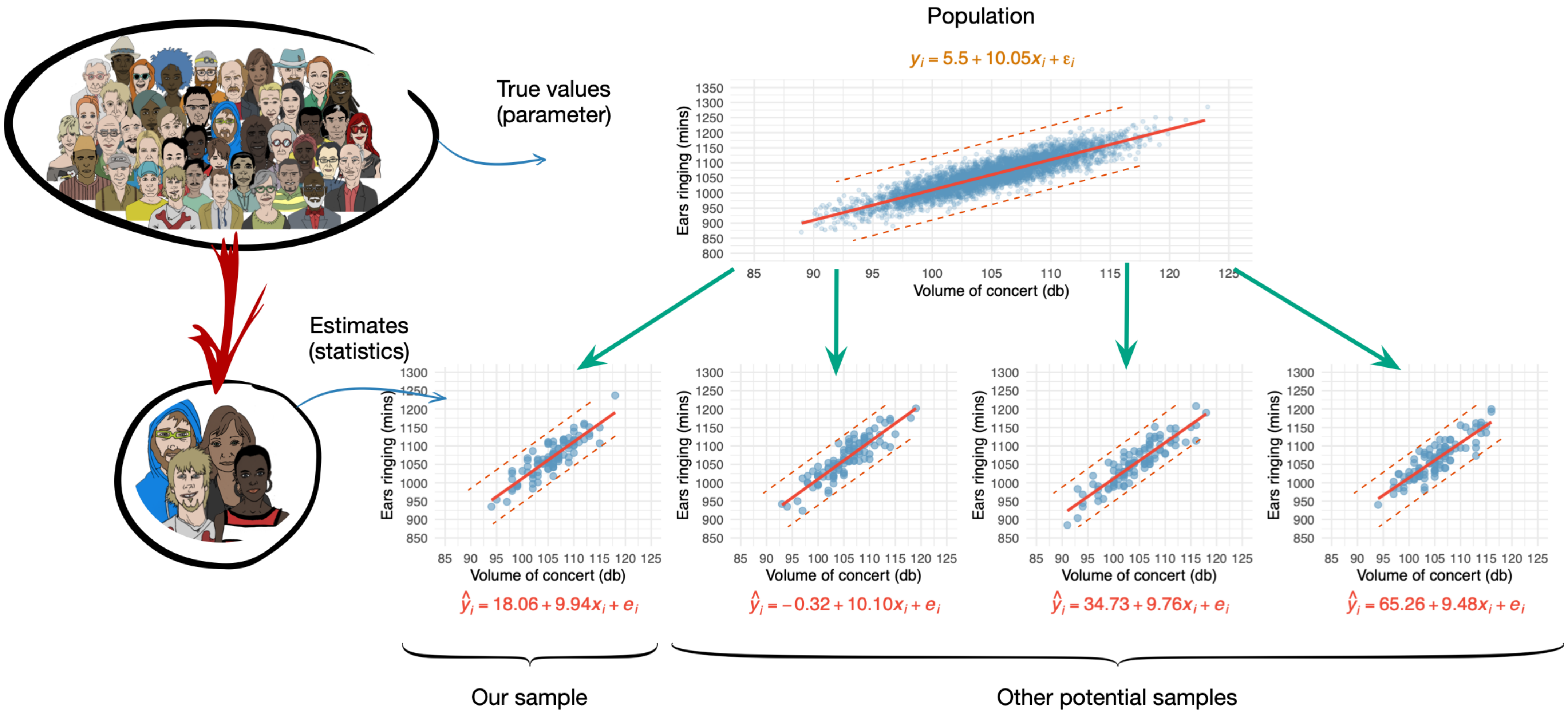
$$\hat{\text{ringing}}_i = \hat{b}_0 + \hat{b}_1 \text{attendance}_i + e_i$$



$$\text{outcome} = \text{model} + \text{error}_i$$



$$\hat{\text{outcome}} = \text{estimated model} + \text{error}_i$$



True values
(parameter)
Population
$y_i = 5.5 + 10.05x_i + \varepsilon_i$
Ears ringing (mins)
Volume of concert (db)
Estimates
(statistics)
$\hat{y}_i = 18.06 + 9.94x_i + e_i$
$\hat{y}_i = -0.32 + 10.10x_i + e_i$
$\hat{y}_i = 34.73 + 9.76x_i + e_i$
$\hat{y}_i = 65.26 + 9.48x_i + e_i$
Our sample
Other potential samples

Least squares estimation: an example

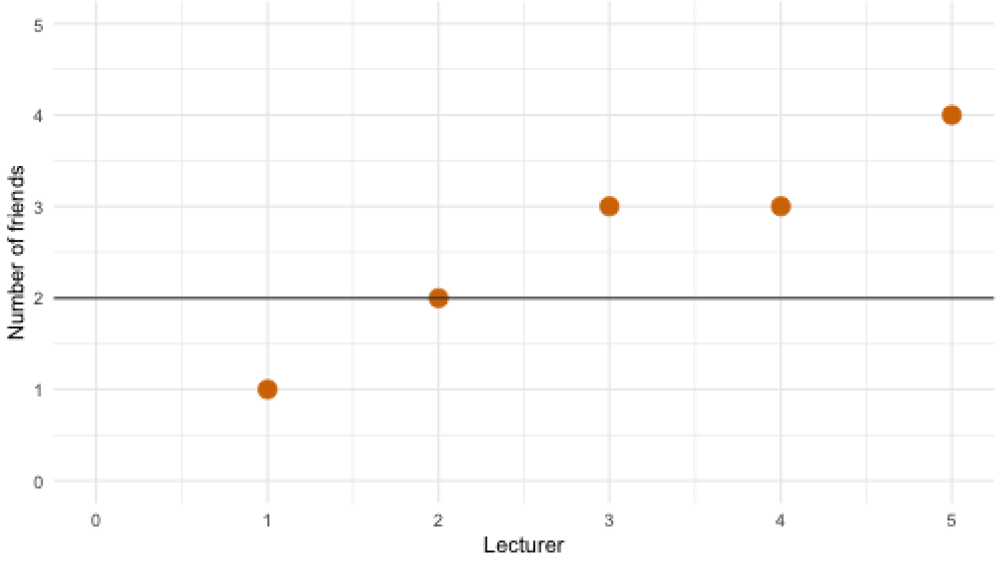
$$\text{friends}_i = \hat{b}_0 + e_i$$

- With no predictors, we predict an outcome from only the intercept $\hat{b}_0$
- In this scenario $\hat{b}_0$ will be the mean value of the outcome
- The model has one predicted value (the mean) for all observations
- Data: 1, 3, 4, 3, 2

Guess 1: $\hat{b}_0$ = 2

$$\text{friends}_i = \hat{b}_0 + \text{error}_i$$
$$\text{error}_i = \text{friends}_i - \hat{b}_0$$

Friends (Y)	Estimate	Error	Squared error
1	2		
2	2		
3	2		
3	2		
4	2		

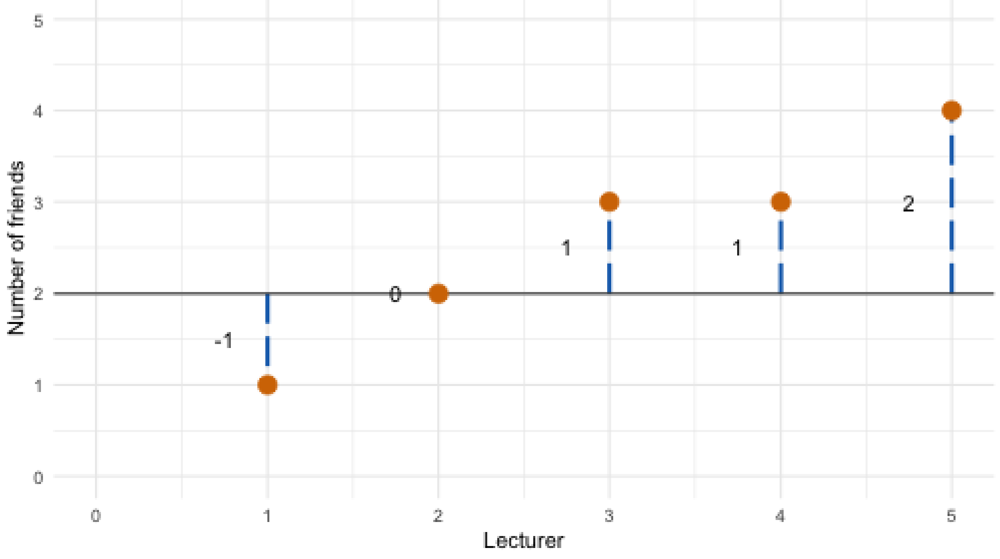


ANDY FIELD

Guess 1: $\hat{b}_0$ = 2

$$\text{friends}_i = \hat{b}_0 + \text{error}_i$$
$$\text{error}_i = \text{friends}_i - \hat{b}_0$$

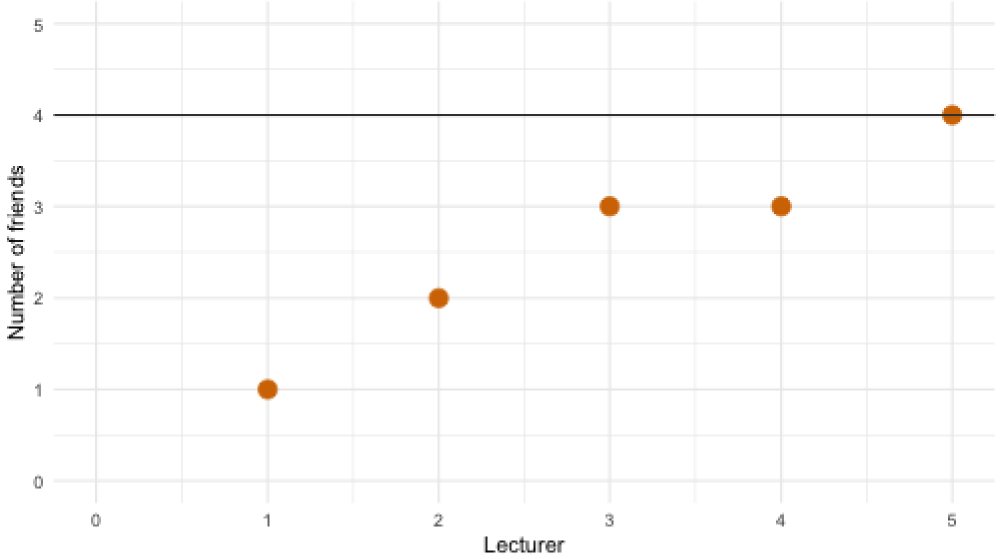
	Friends (Y)	Estimate	Error	Squared error
	1	2	-1	1
	2	2	0	0
	3	2	1	1
	3	2	1	1
	4	2	2	4
Total	—	—	—	7.00



Guess 2: $\hat{b}_0$ = 4

$$\text{friends}_i = \hat{b}_0 + \text{error}_i$$
$$\text{error}_i = \text{friends}_i - \hat{b}_0$$

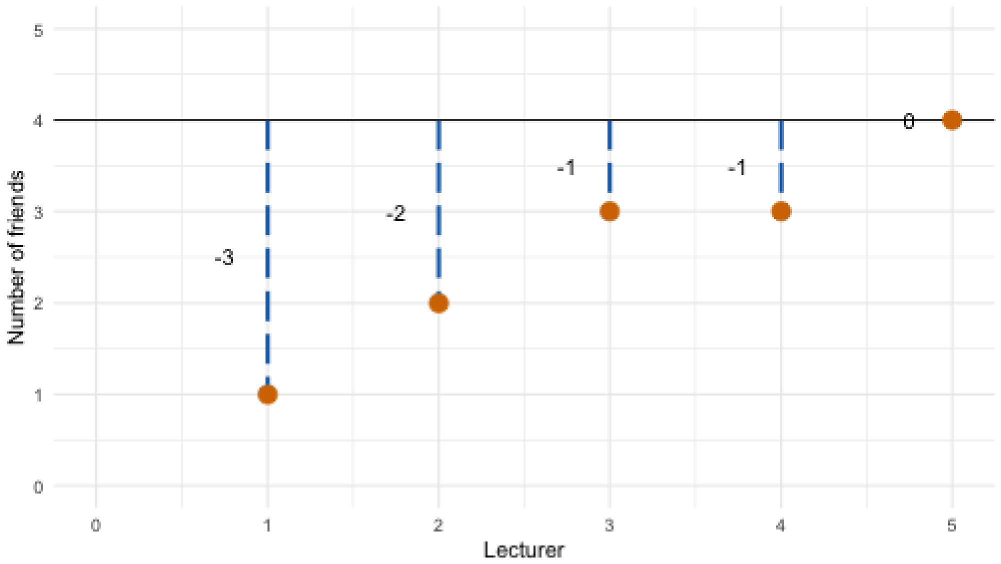
Friends (Y)	Estimate	Error	Squared error
1	4		
2	4		
3	4		
3	4		
4	4		

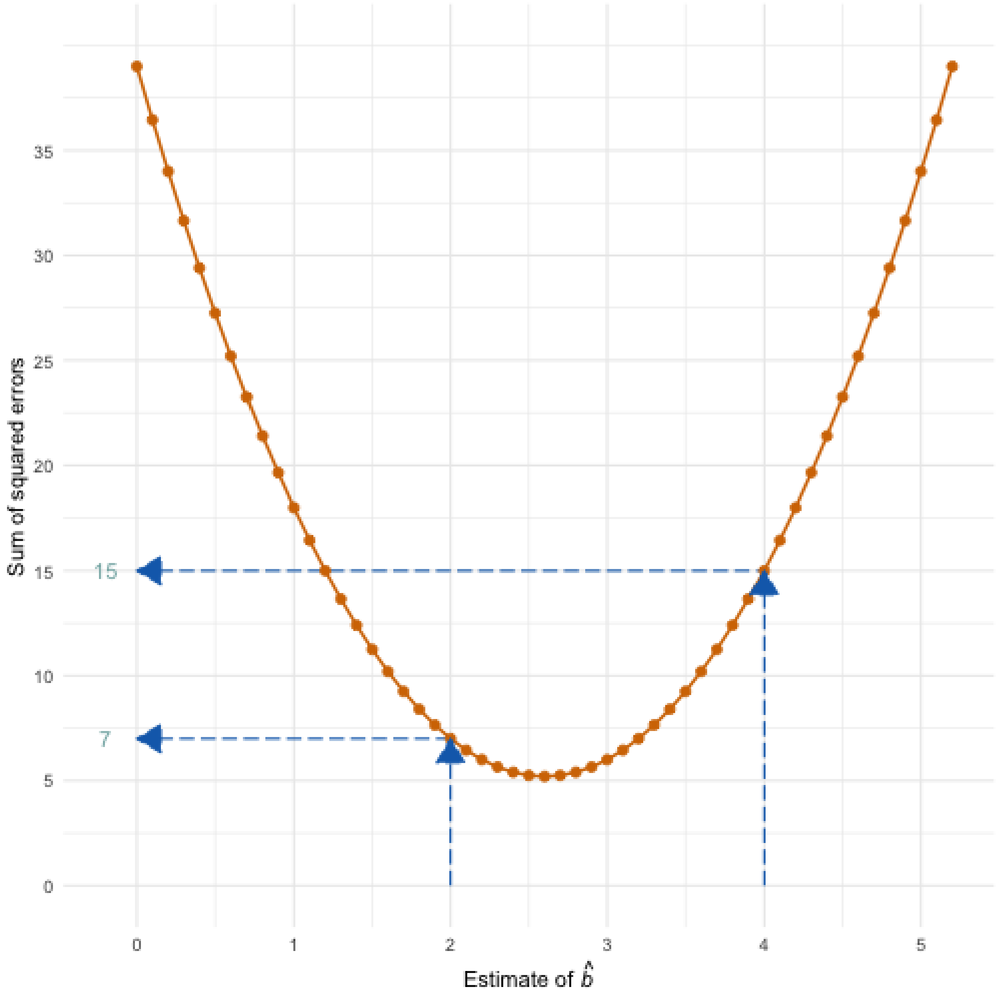


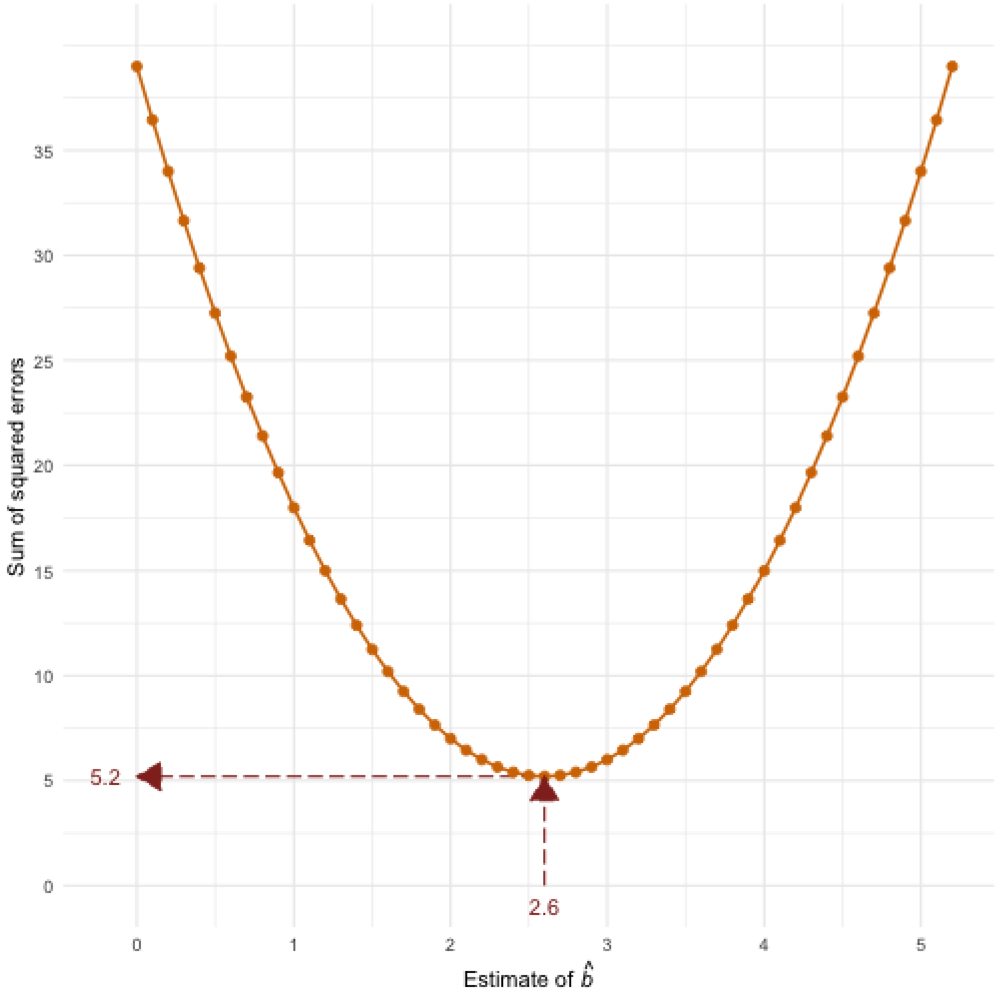
Guess 2: $\hat{b}_0$ = 4

$$\text{friends}_i = \hat{b}_0 + \text{error}_i$$
$$\text{error}_i = \text{friends}_i - \hat{b}_0$$

	Friends (Y)	Estimate	Error	Squared error
	1	4	-3	9
	2	4	-2	4
	3	4	-1	1
	3	4	-1	1
	4	4	0	0
Total	—	—	—	15.00




Sum of squared errors
Estimate of $\hat{b}$
15
7
0
5
10
15
20
25
30
35
0
1
2
3
4
5

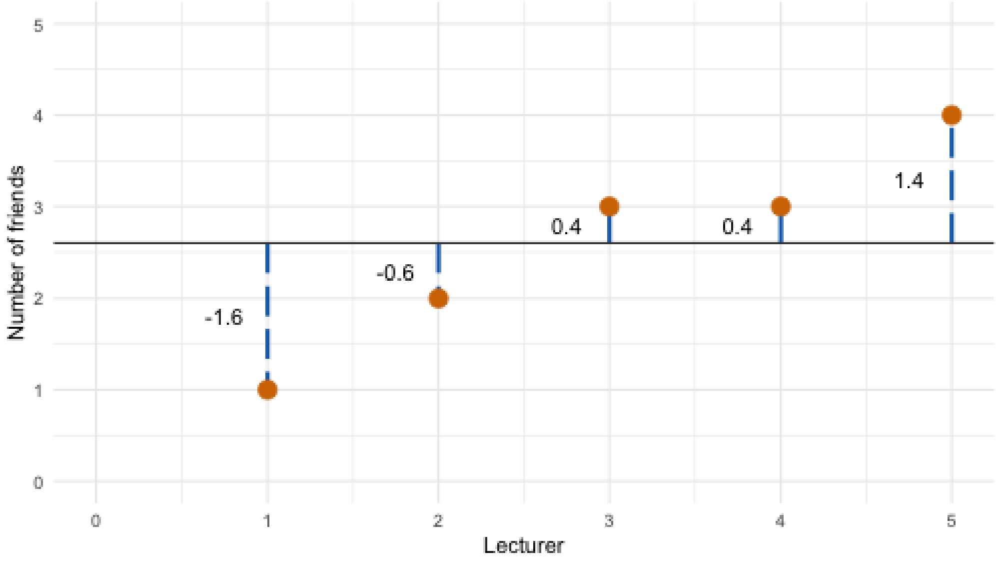


Sum of squared errors
35
30
25
20
15
10
5
0
5.2
2.6
0
1
2
3
4
5
Estimate of $\hat{b}$

OLS estimate: $\hat{b}_0$ = 2.6

$$\text{friends}_i = \hat{b}_0 + \text{error}_i$$
$$\text{error}_i = \text{friends}_i - \hat{b}_0$$

	Friends (Y)	Estimate	Error	Squared error
	1	2.6	-1.6	2.56
	2	2.6	-0.6	0.36
	3	2.6	0.4	0.16
	3	2.6	0.4	0.16
	4	2.6	1.4	1.96
Total	—	—	—	5.20



- Data: 1, 3, 4, 3, 2

$$\widehat{\text{friends}} = \frac{\sum_{i=1}^{n} x_i}{n}$$

- Add up the scores

$$\begin{aligned}\sum_{i=1}^{n} x_i &= 1 + 3 + 4 + 3 + 2 \\ &= 13\end{aligned}$$

- Divide by the number of scores, *n*

$$\begin{aligned}\frac{\sum_{i=1}^{n} x_i}{n} &= \frac{13}{5} \\ &= 2.6\end{aligned}$$

Summary

- People try to make statistics seem complicated but boils down to simple ideas:
    - We predict one variable from a model containing one or more predictors
    - There is always error in prediction
    - The model you fit represents hypotheses
    - The model you fit is typically a variation on the the linear model (GLM)
    - If you understand this one model, you understand most of psychological statistics
- Parameter estimates (*b*)
    - Tell us about the shape of the model
    - Tell us about size and direction of relationship between predictor(s) and outcome
    - They are estimated from the sample data
- Estimation
    - Ordinary least squares (OLS) estimation is one (of many) methods for estimating parameters
    - The mean is an example of an OLS estimator