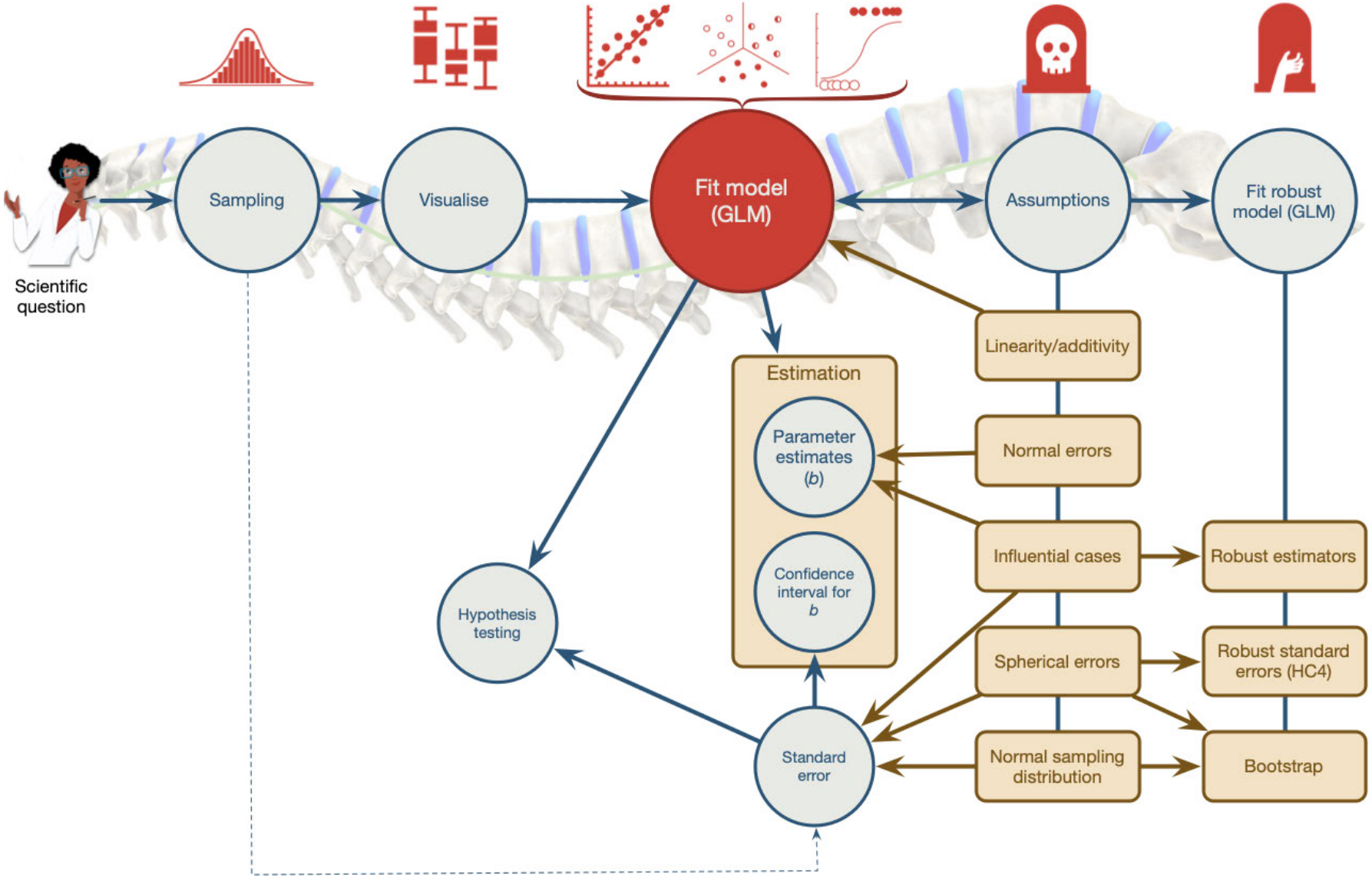
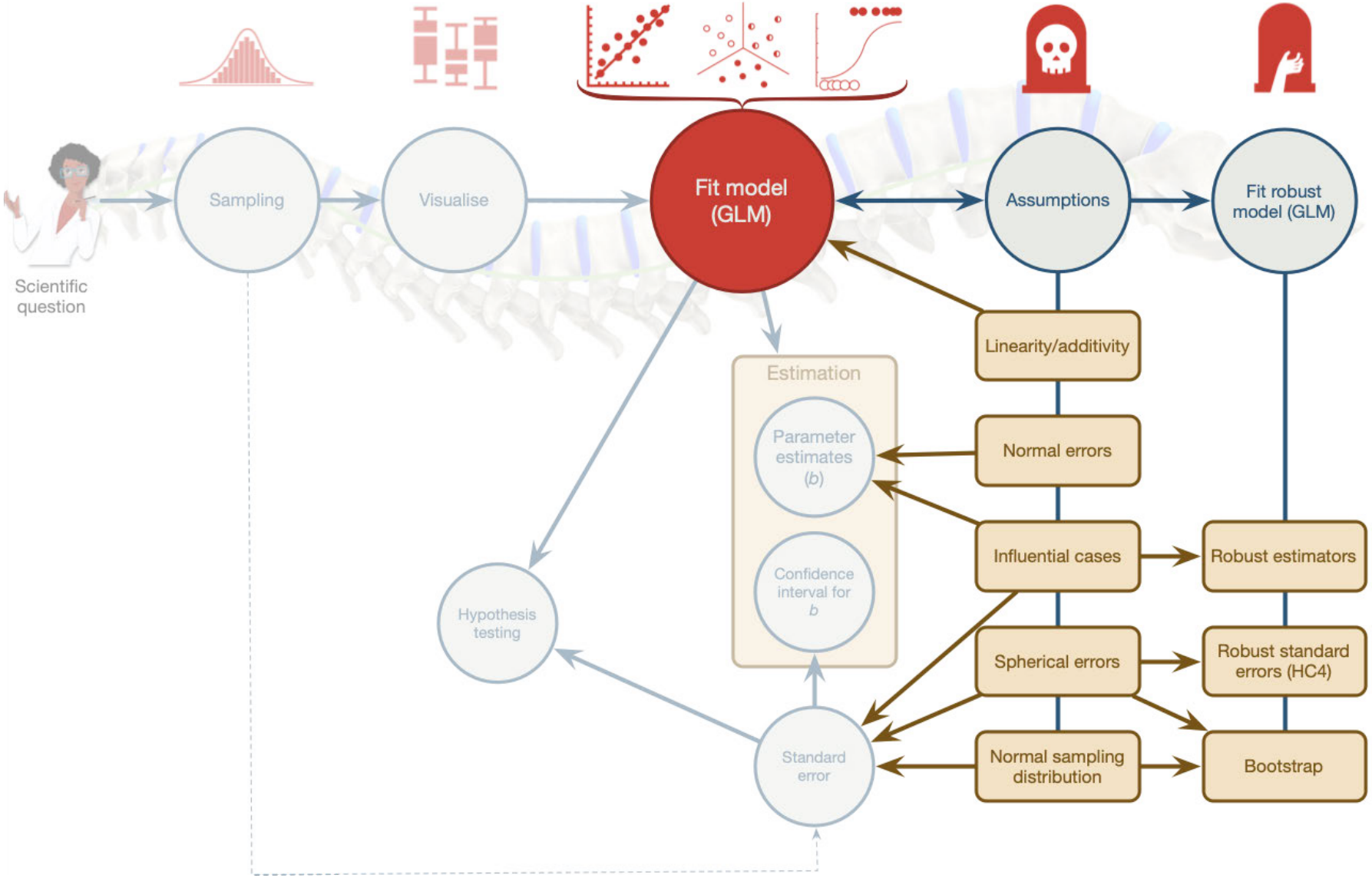

The Beast of Bias
Professor Andy Field
@profandyfield
www.youtube.com/user/ProfAndyField/
www.discoveringstatistics.com
www.milton-the-cat.rocks
www.discovr.rocks


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


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

Learning outcomes

Describe sources of bias in the General Linear Model

- Outliers
- Linearity and additivity
- Spherical residuals
  - Homoscedastic errors
  - Independent errors
- Normality of something-or-other
  - Residuals
  - Sampling distribution
- Robust methods
  - Bootstrapping

Recap: The General Linear Model (GLM)

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

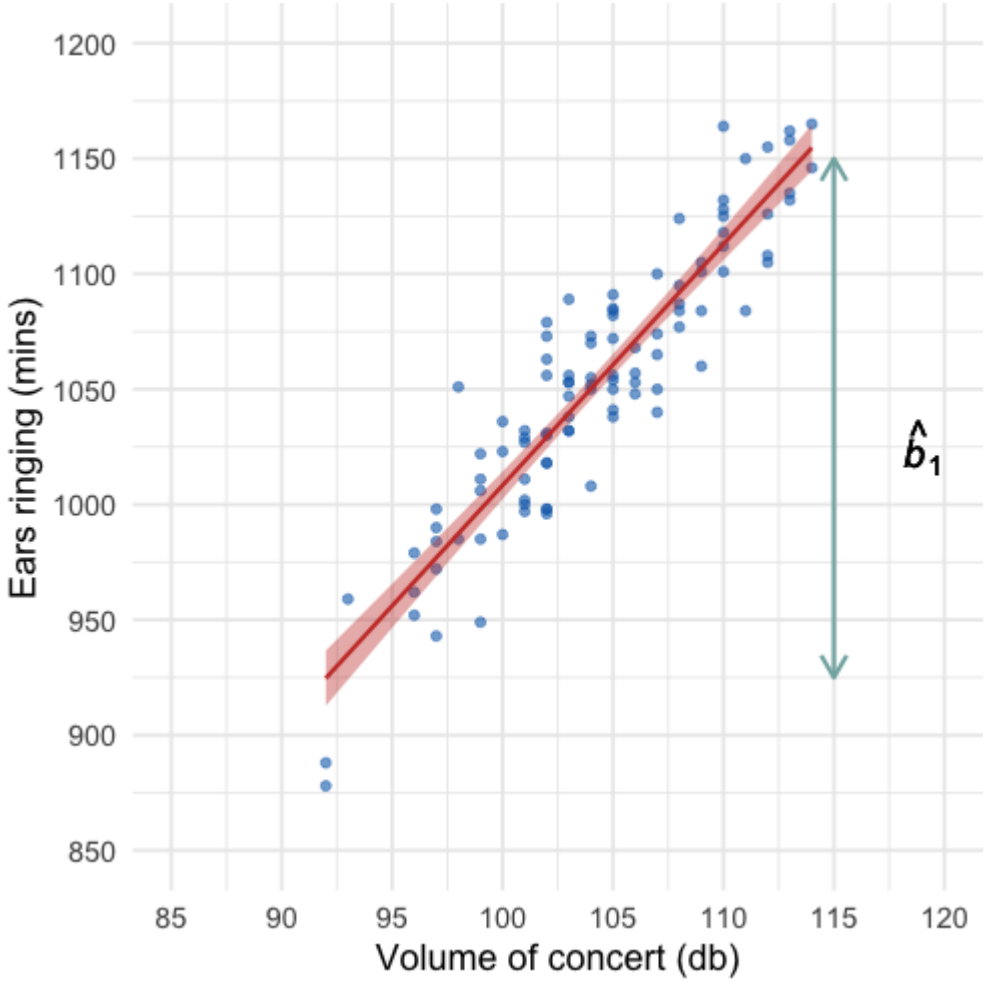
$\hat{b}_n$

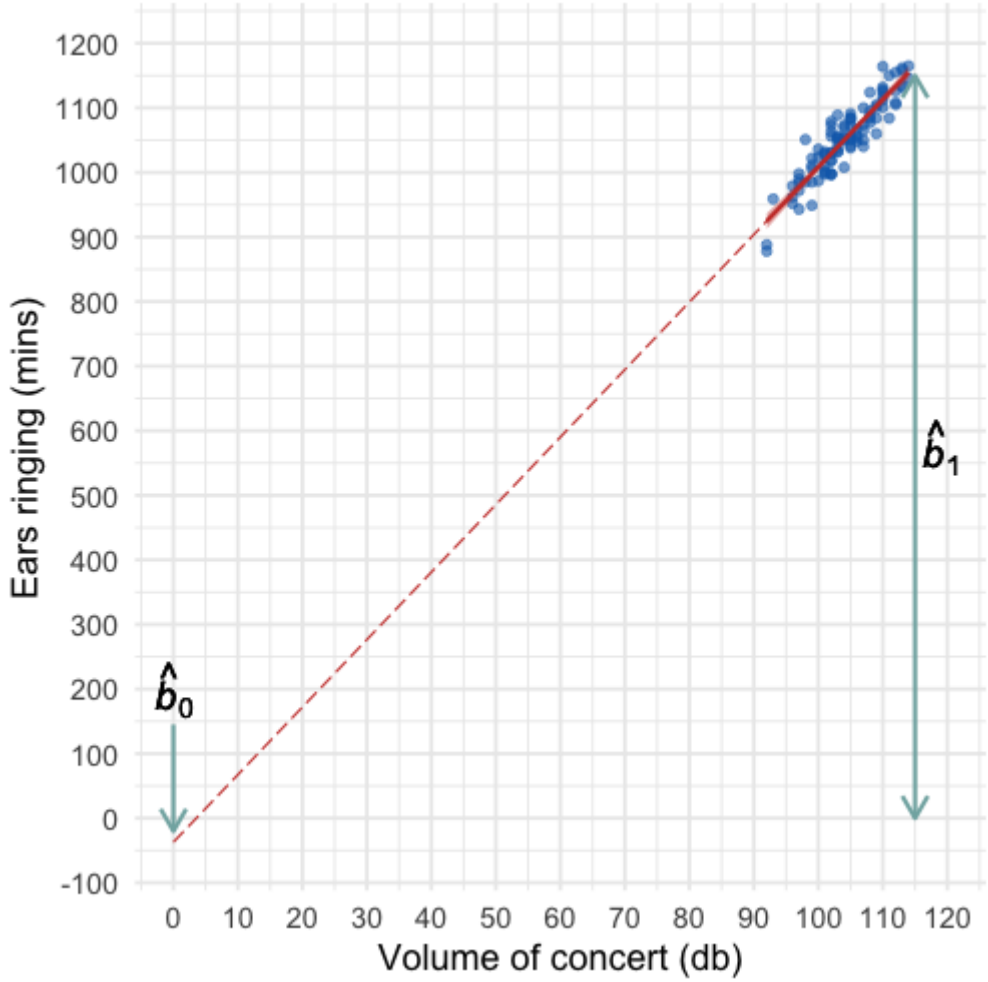
- Estimate of parameter for a predictor, *n*
  - Direction/strength of relationship/effect
  - Difference in means

$\hat{b}_0$

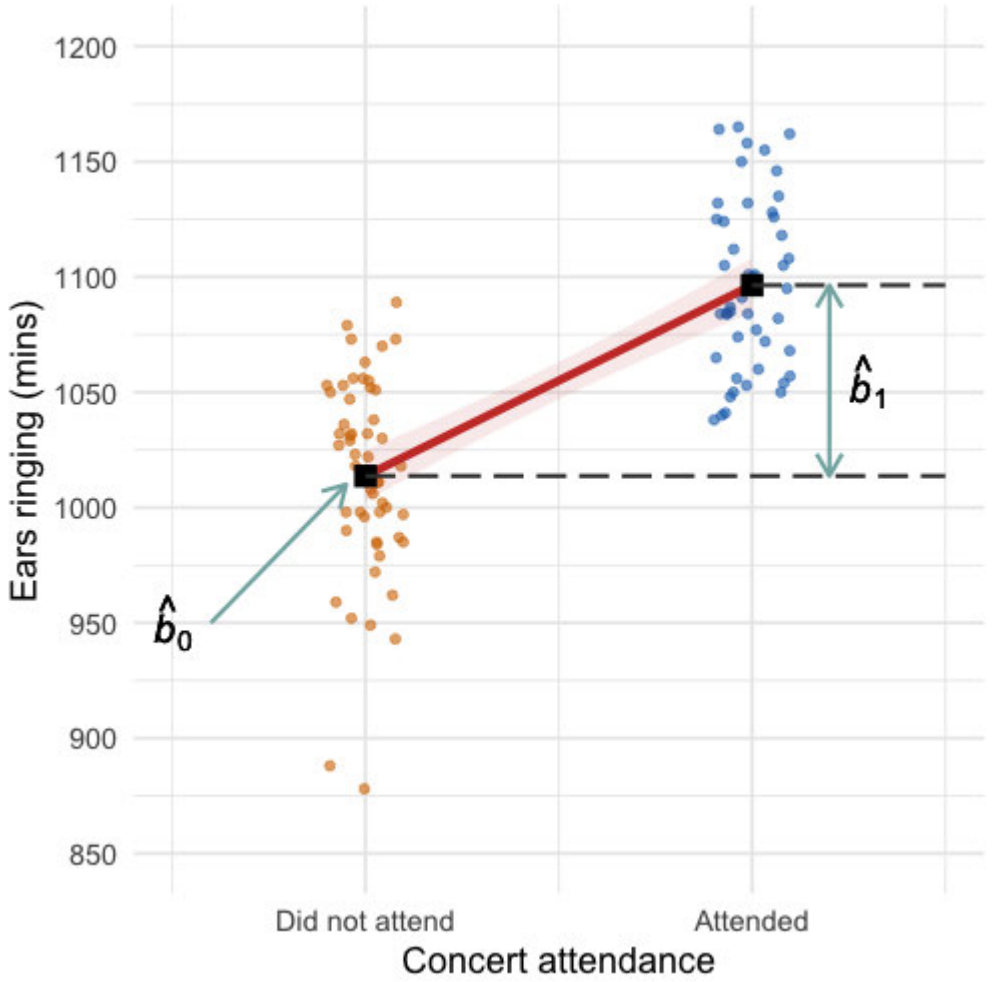
- Estimate of the value of the outcome when predictor(s) = 0 (intercept)

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

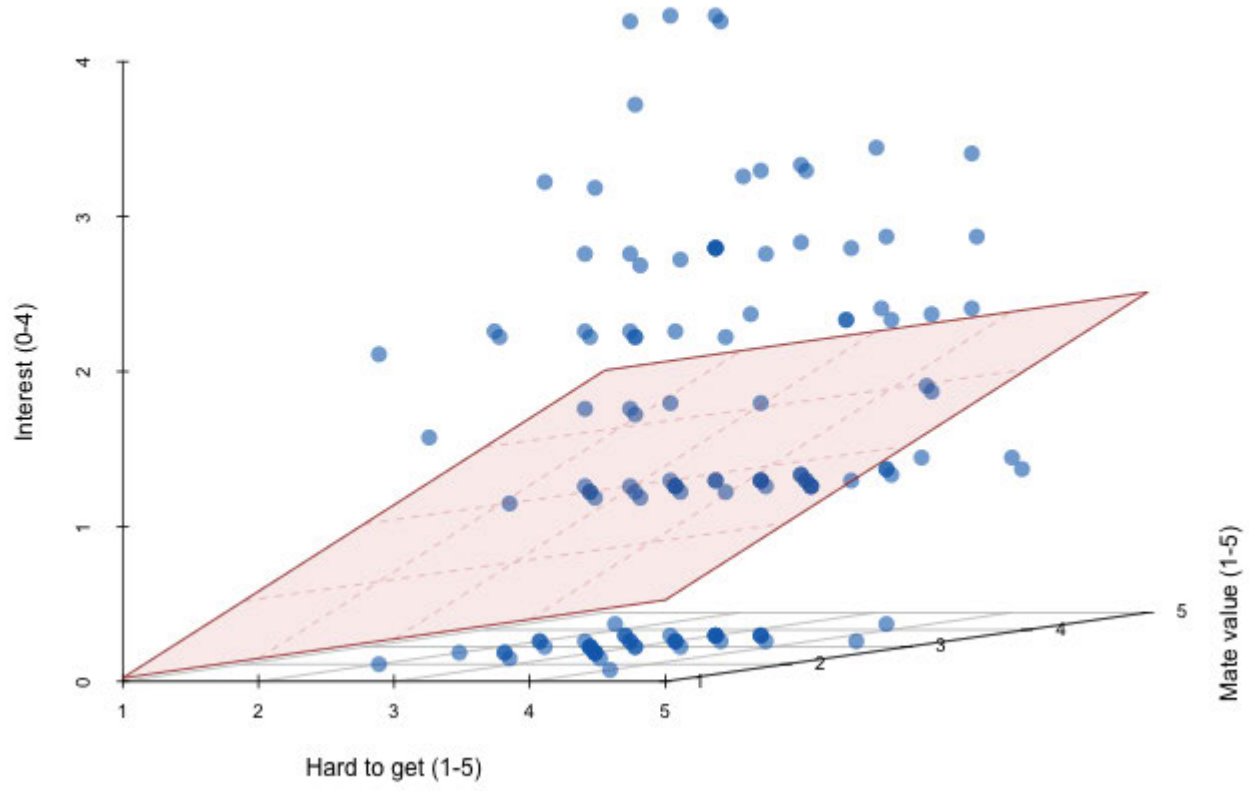




$$\widehat{\text{ringing}}_i = \hat{b}_0 + \hat{b}_1 \text{attendance}_i + e_i$$



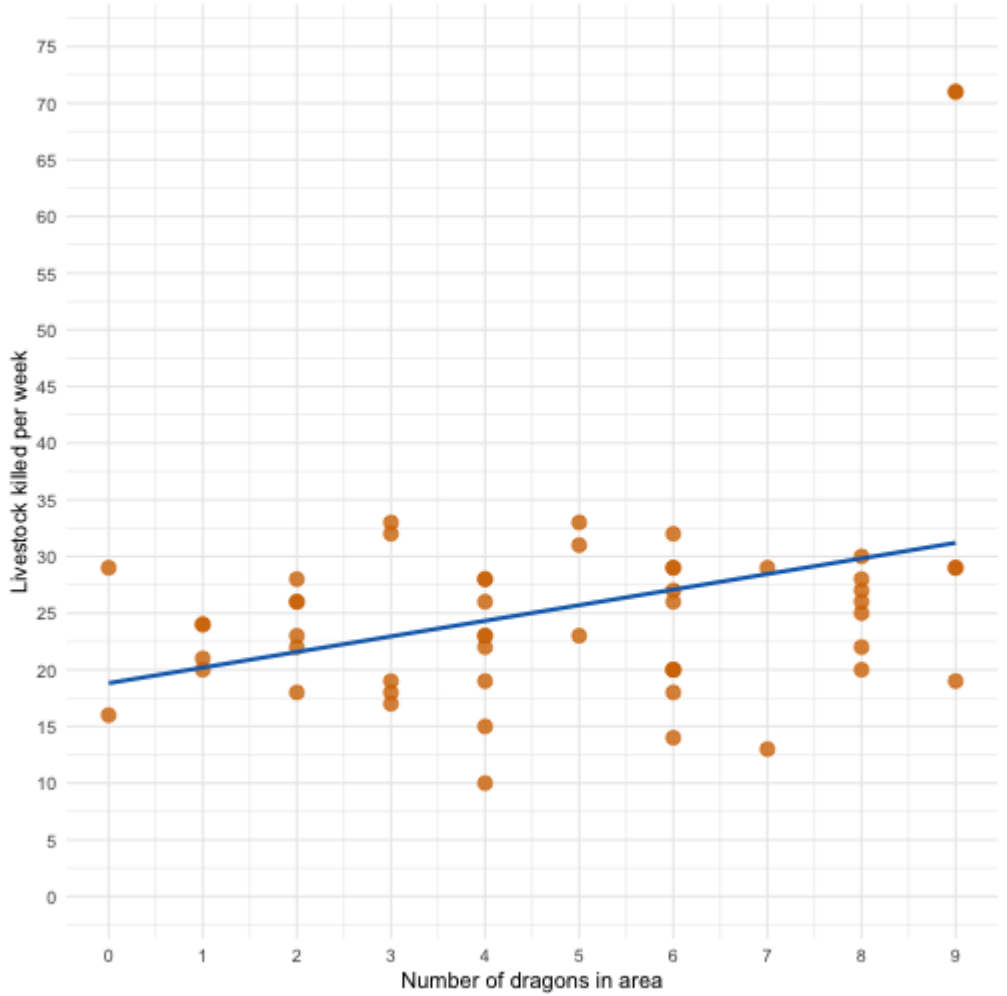
$$\hat{\text{interest}}_i = \hat{b}_0 + \hat{b}_1 \text{hard to get}_i + \hat{b}_2 \text{mate value}_i + e_i$$

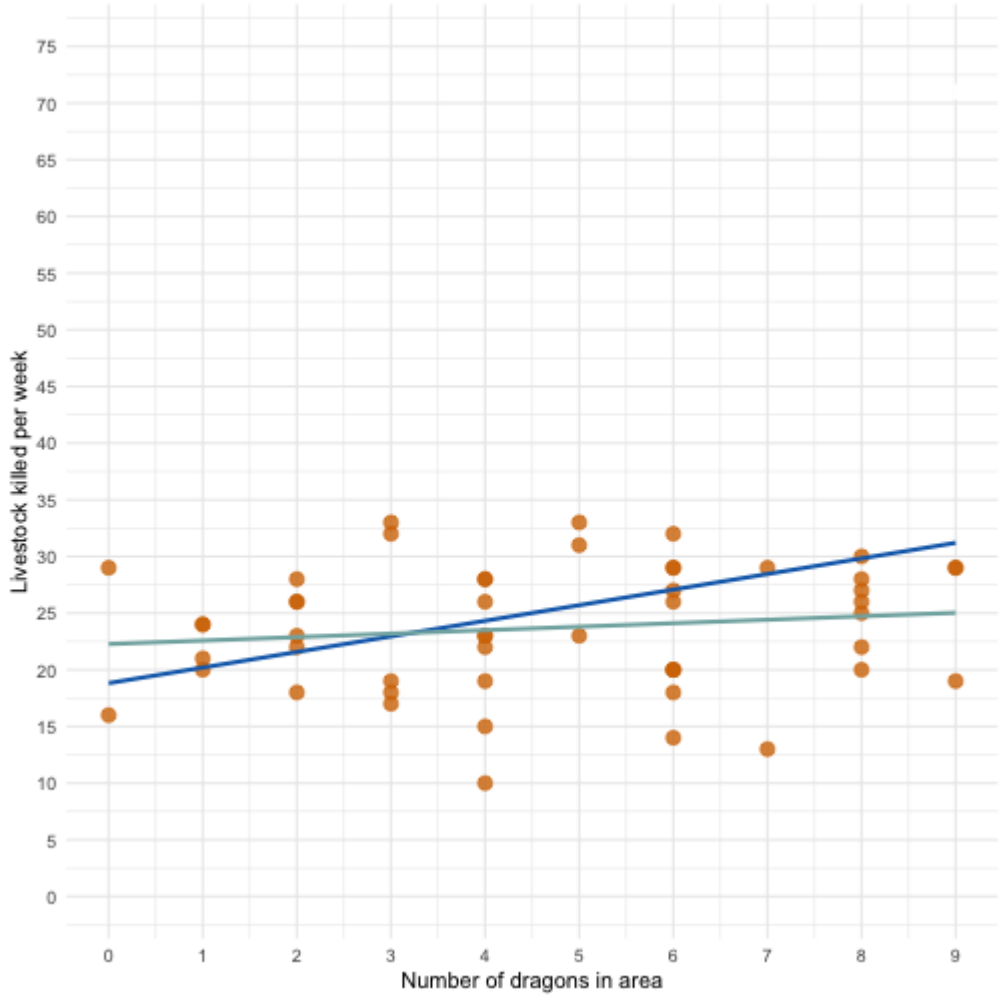


Part 1: Outliers and do dragons eat sheep?

"Coz they eat all our sheep"

Sir Knight Zach, Defender of the world of Military


Livestock killed per week
75
70
65
60
55
50
45
40
35
30
25
20
15
10
5
0
0
1
2
3
4
5
6
7
8
9
Number of dragons in area

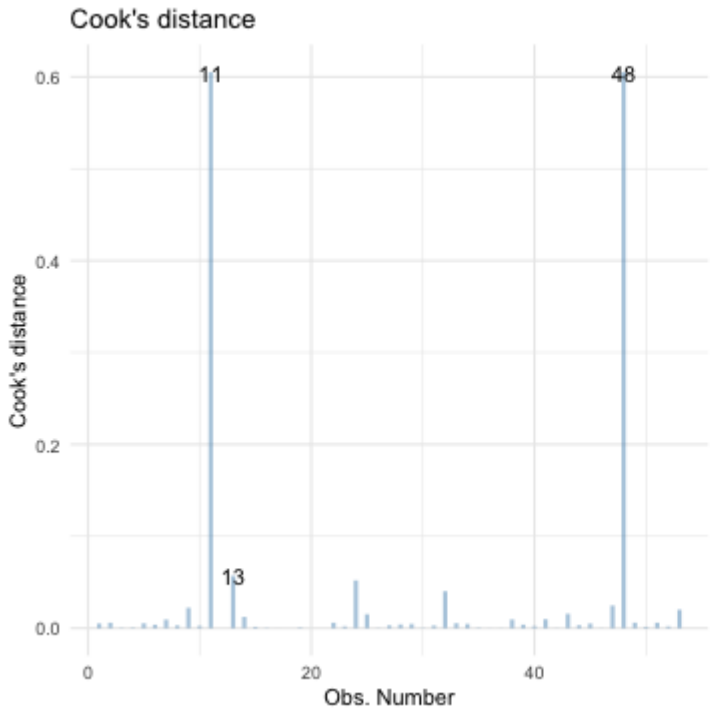

Livestock killed per week
75
70
65
60
55
50
45
40
35
30
25
20
15
10
5
0
0
1
2
3
4
5
6
7
8
9
Number of dragons in area

Detecting outliers and influential cases

- Graphs
  - Scatterplots (less helpful with several predictors)
  - Histograms
- Standardized residual
  - In an average sample, 95% of standardized residuals should lie between ± 2
  - 99% of standardized residuals should lie between ± 2.5
  - Any case for which the absolute value of the standardized residual is 3 or more, is likely to be an outlier
- Cook's distance
  - Measures the *influence* of a single case on the model as a whole
  - Absolute values greater than 1 may be cause for concern (Cook & Weisberg , 1982)
- DF beta statistics (unstandardized or standardized)
  - The change in *b* when a case is removed
  - Be wary of standardized values with absolute values > 1

Influential cases

```
out_lm <- lm(livestock ~ dragons, data = out_tib)
# plot(out_lm,  which = 4)
ggplot2::autoplot(out_lm,which = 4,colour = "#5c97bf", alpha = 0.5, size = 1) +
  theme_minimal()
```



Full sample

term	estimate	std.error	statistic	p.value
(Intercept)	18.82	2.93	6.43	0.00
dragons	1.38	0.53	2.59	0.01

Influential cases removed

term	estimate	std.error	statistic	p.value
(Intercept)	22.27	1.64	13.60	0.00
dragons	0.31	0.31	0.99	0.33

Robust estimation (optional)

Normal model (OLS)

```
out_lm <- lm(livestock ~ dragons,
  data = out_tib)
broom::tidy(out_lm)
```

term	estimate	std.error	statistic	p.value
(Intercept)	18.82	2.93	6.43	0.00
dragons	1.38	0.53	2.59	0.01

Robust model

```
out_rob <- robust::lmRob(livestock ~ dragons,
  data = out_tib)
summary(out_rob)
```

	Estimate	Std. Error	t value	Pr(>|t|)
(Intercept)	22.35	1.81	12.36	0.00
dragons	0.30	0.34	0.88	0.38

Part 2: Linearity, spherical errors and do dragons kidnap royalty?

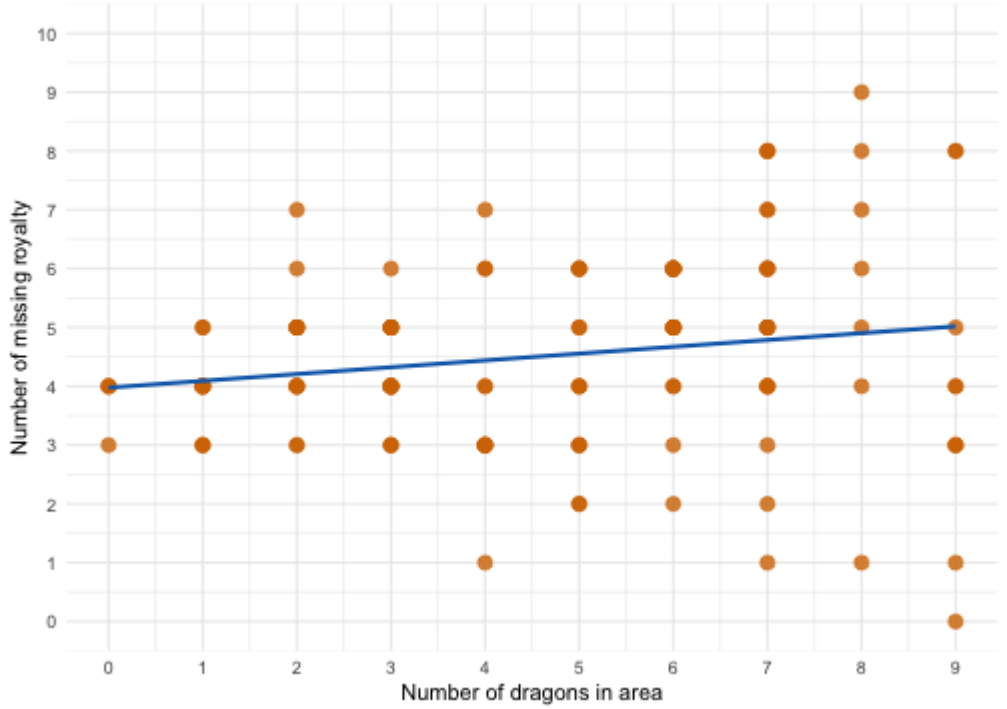
"Coz they kidnap the princesses"

Sir Knight Zach, Defender of the world of Military

Are more princesses kidnapped in areas with more dragons?

$$\hat{\text{royalty}}_i = 3.98 + 0.12\text{dragons}_i + e_i$$

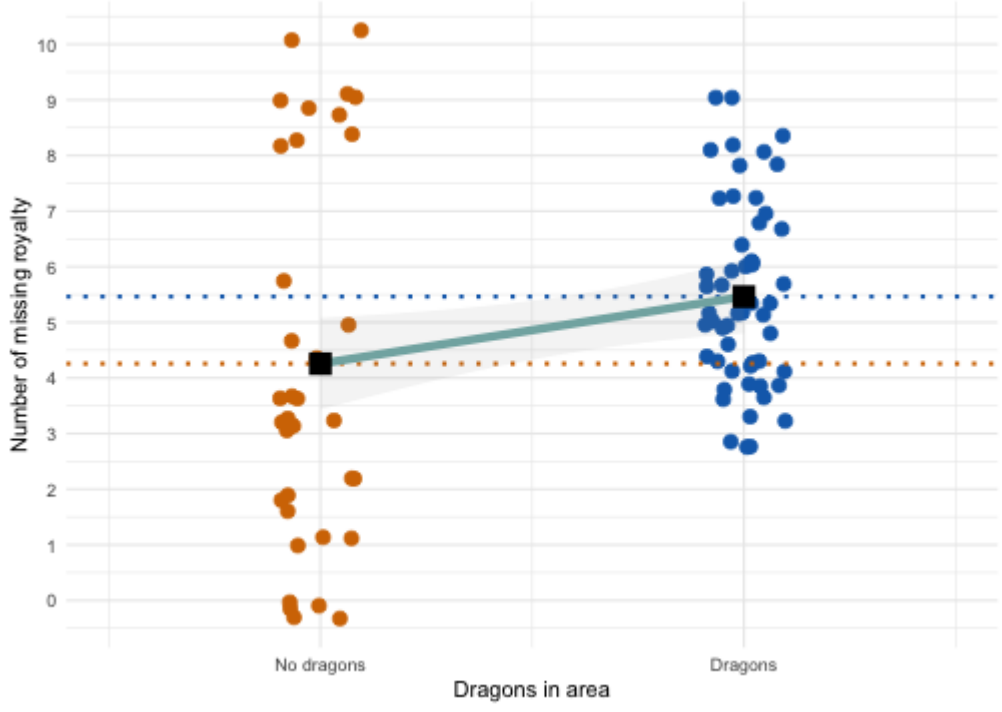
term	estimate	std.error	statistic	p.value	conf.low	conf.high
(Intercept)	3.98	0.30	13.32	0.00	3.39	4.57
dragons	0.12	0.06	2.07	0.04	0.01	0.23



Are more princesses kidnapped in areas that have dragons?

$$\text{royalty}_i = 4.26 + 1.21\text{dragons}_i + e_i$$

term	estimate	std.error	statistic	p.value	conf.low	conf.high
(Intercept)	4.26	0.41	10.28	0.00	3.43	5.08
dragonsDragons	1.21	0.53	2.27	0.03	0.15	2.26



Key assumptions of the General Linear Model

Linearity and additivity

Spherical errors

The population model should have:

- Homoscedastic errors
  - Inspect the model residuals
- Independent errors
  - Inspect the model residuals

Normality of something-or-other

- Population model errors
- Sampling distribution

Linearity and additivity

The relationship between predictor(s) and outcome is, in reality, linear

$$\hat{\text{royalty}}_i = \hat{b}_0 + \hat{b}_1 \text{dragons}_i + e_i$$

The combined effect of predictors is additive

$$\hat{\text{royalty}}_i = \hat{b}_0 + \hat{b}_1 \text{dragons}_i + \hat{b}_2 \text{strict regime}_i + e_i$$

Test linearity using plots. If the data cloud looks banana shaped (curved), linearity probably can't be assumed.

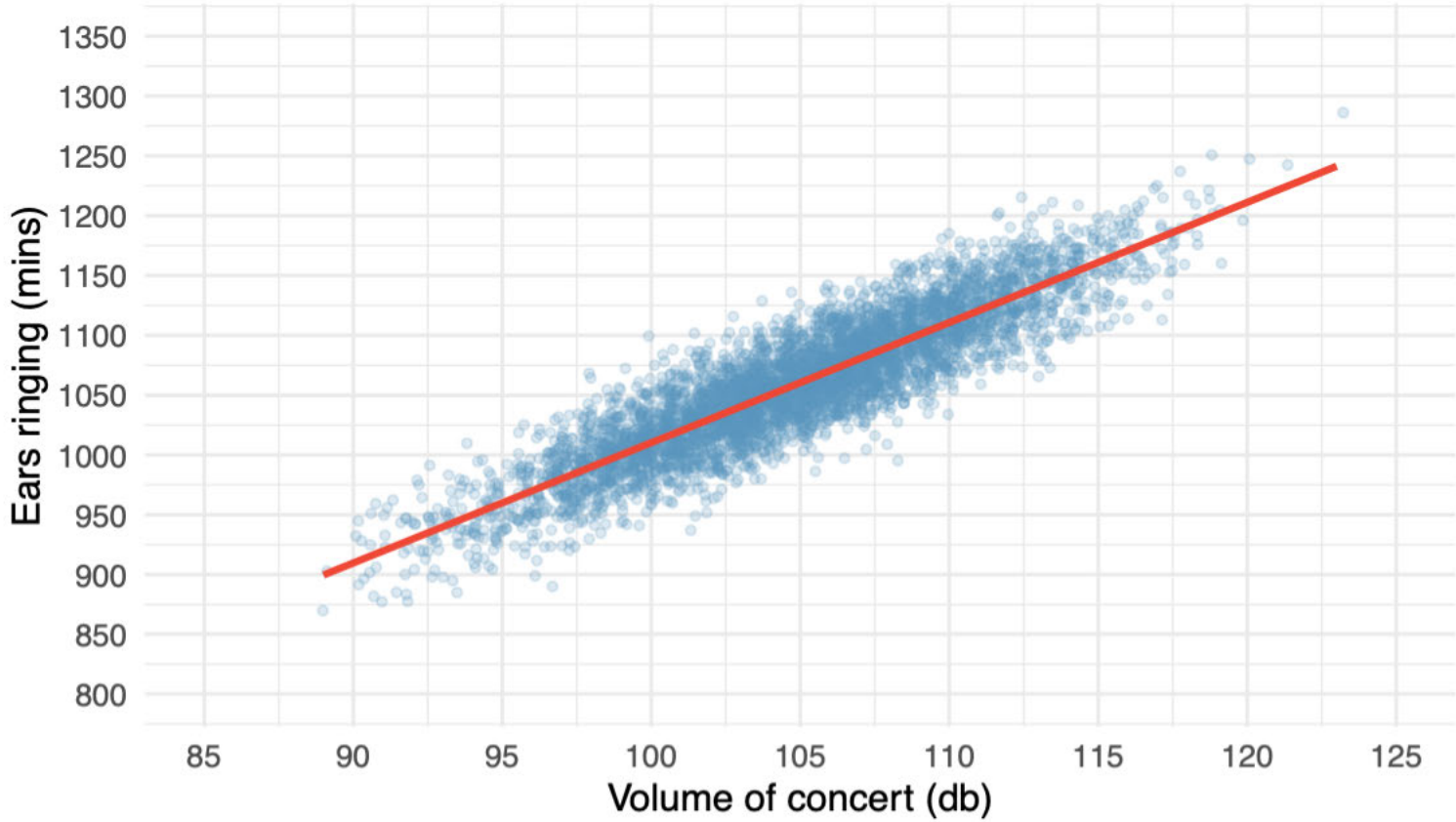
Errors vs. Residuals

- A model's ERRORs refer to the differences between predicted values and observed values of the outcome variable in the population model
- These values cannot be observed

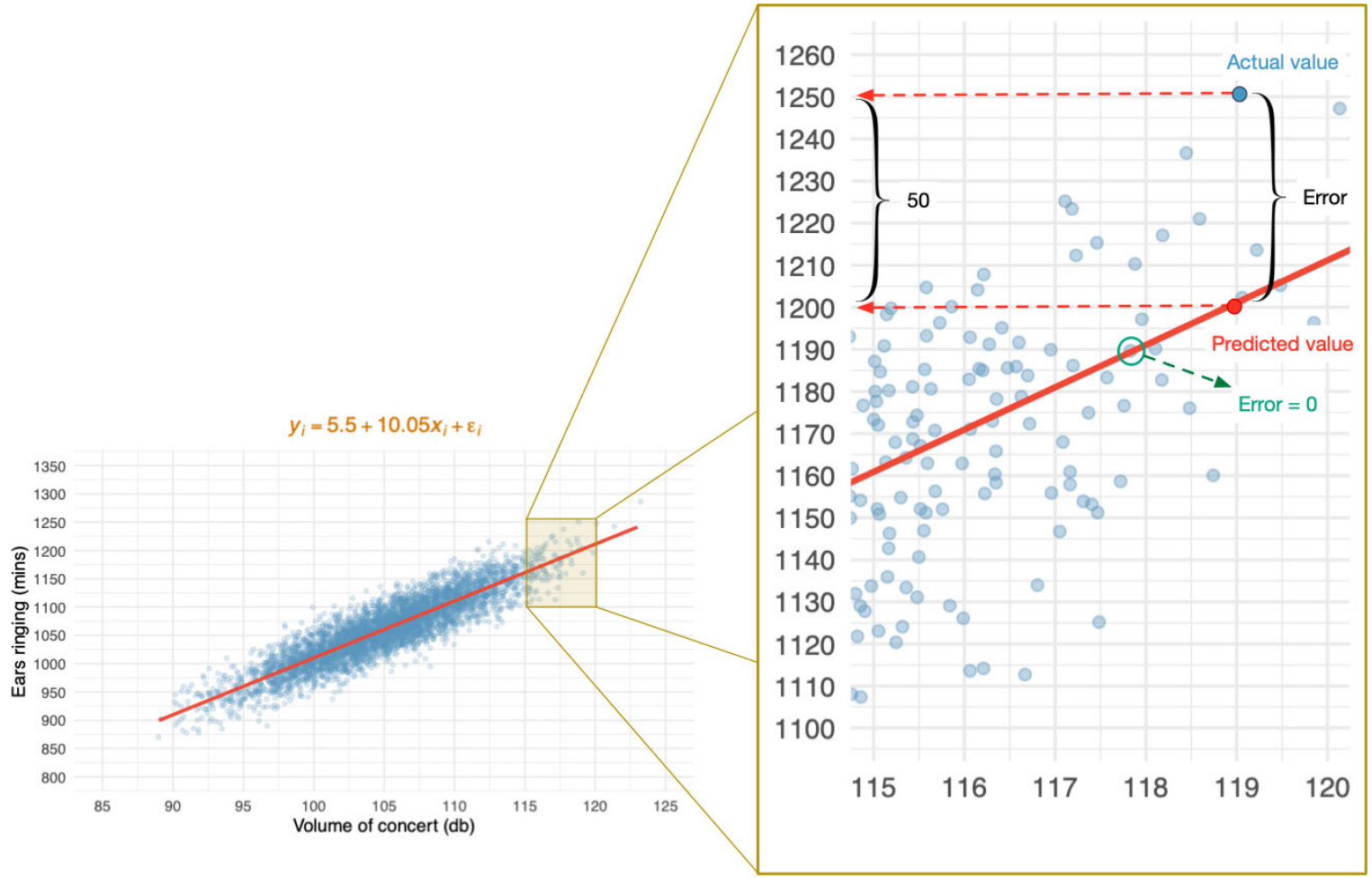
- A model's RESIDUALs refer to the differences between predicted values and observed values of the outcome variable in the sample model
- These values can be observed and are representative of the population model errors.

Errors vs residuals

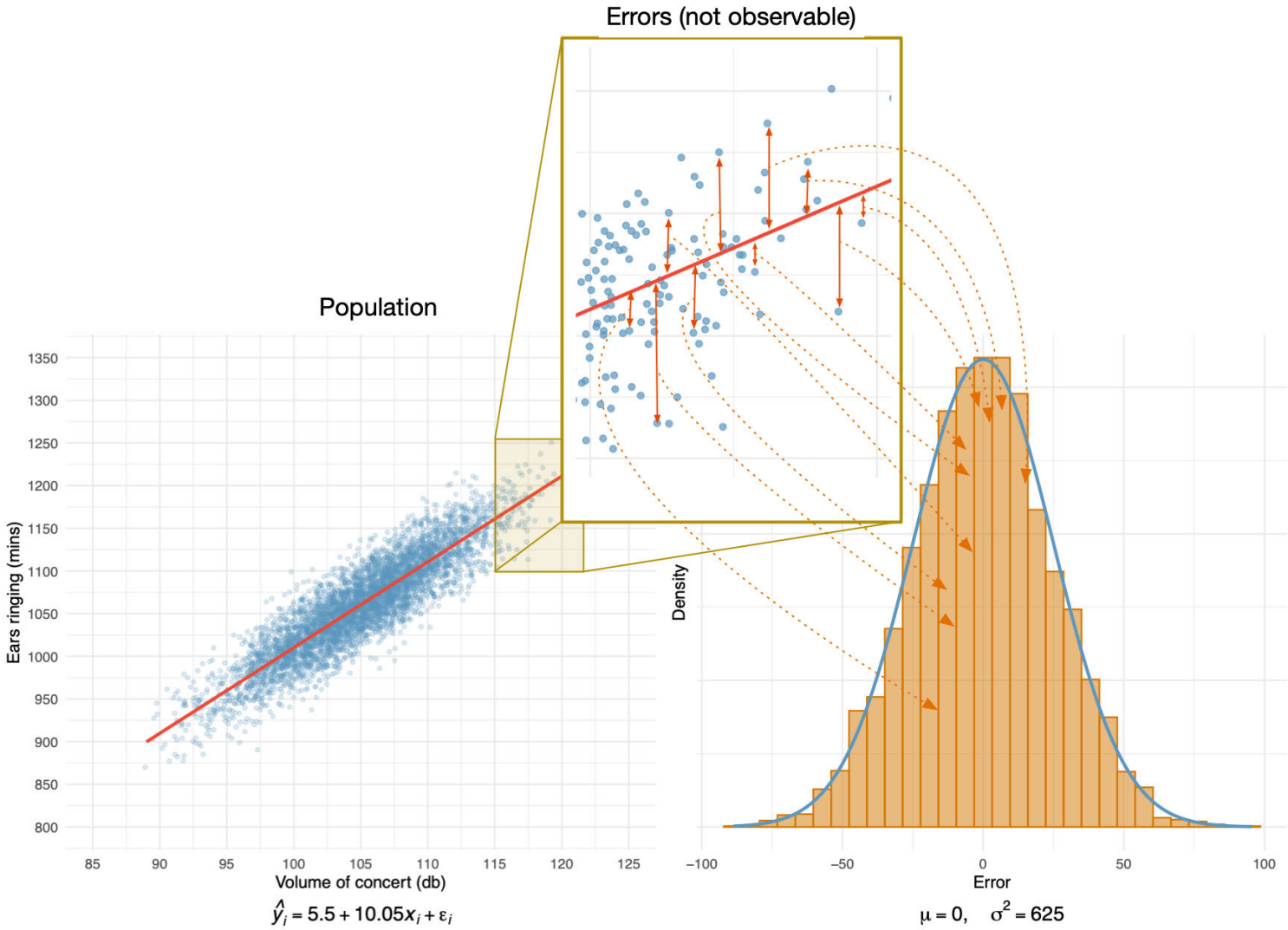
$$y_i = 5.5 + 10.05x_i + \varepsilon_i$$



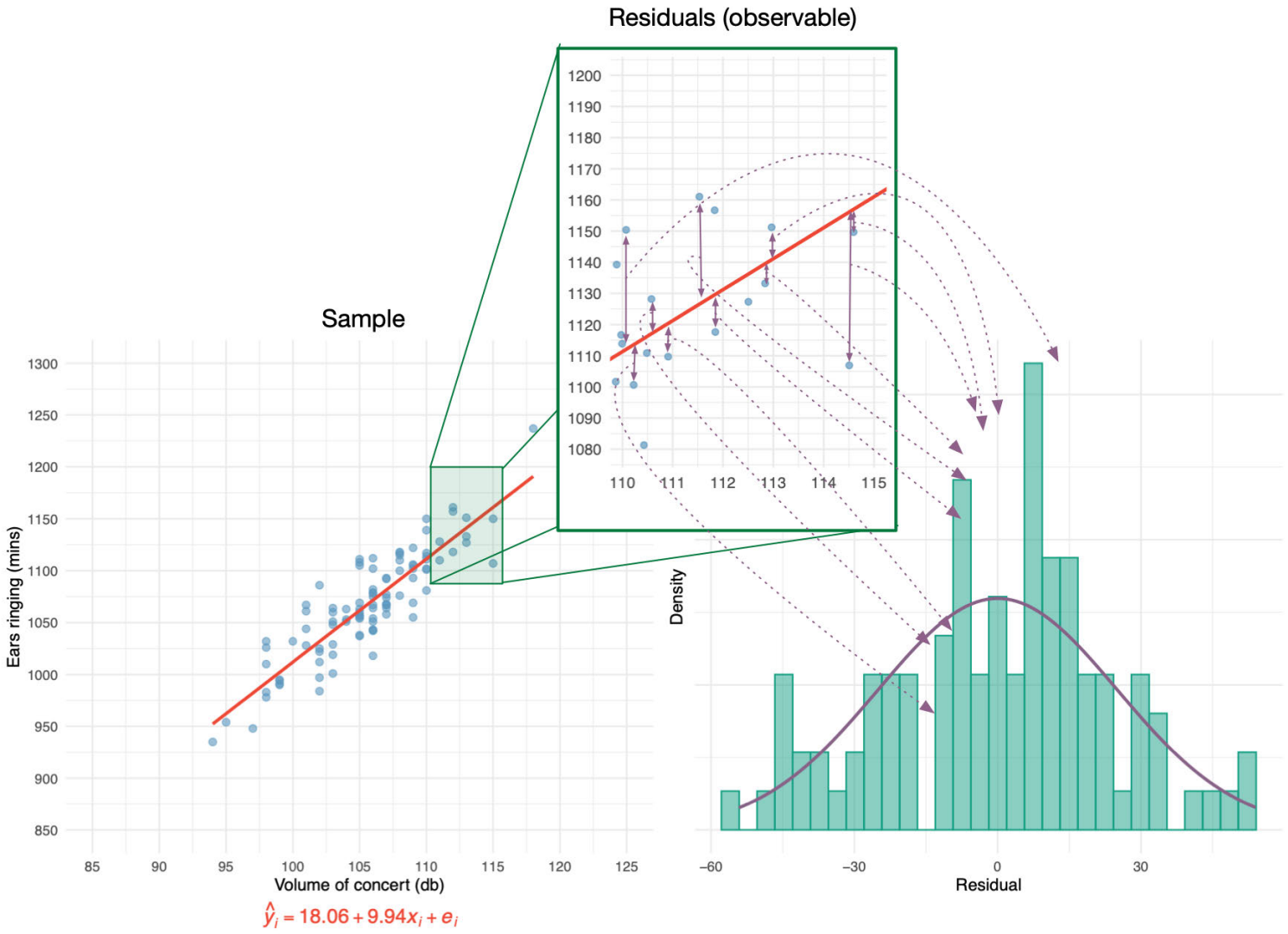
Errors (Population model)



Errors (not observable)



Residuals (are observable)



Spherical errors

Errors should be independent

- The population error in prediction for one case should not be related to the error in prediction for another case (autocorrelation).
- Independent observations tend to lead to independent errors
- Because we cannot observe population errors we inspect the sample residuals

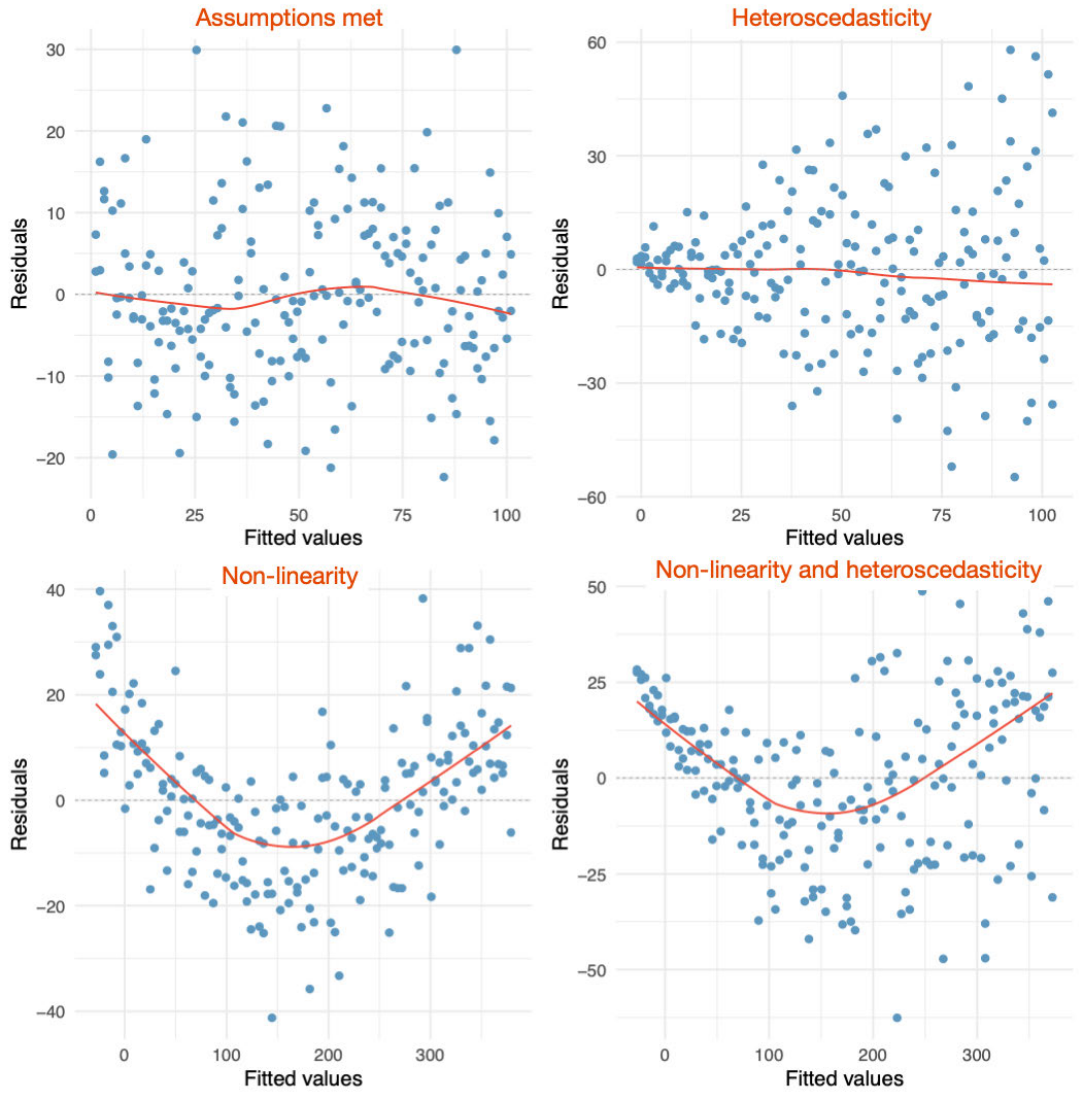
Errors should be homoscedastic

- Variance of population errors (residuals) should be consistent at different values of the predictor variable
- Because we cannot observe population errors we inspect the sample residuals

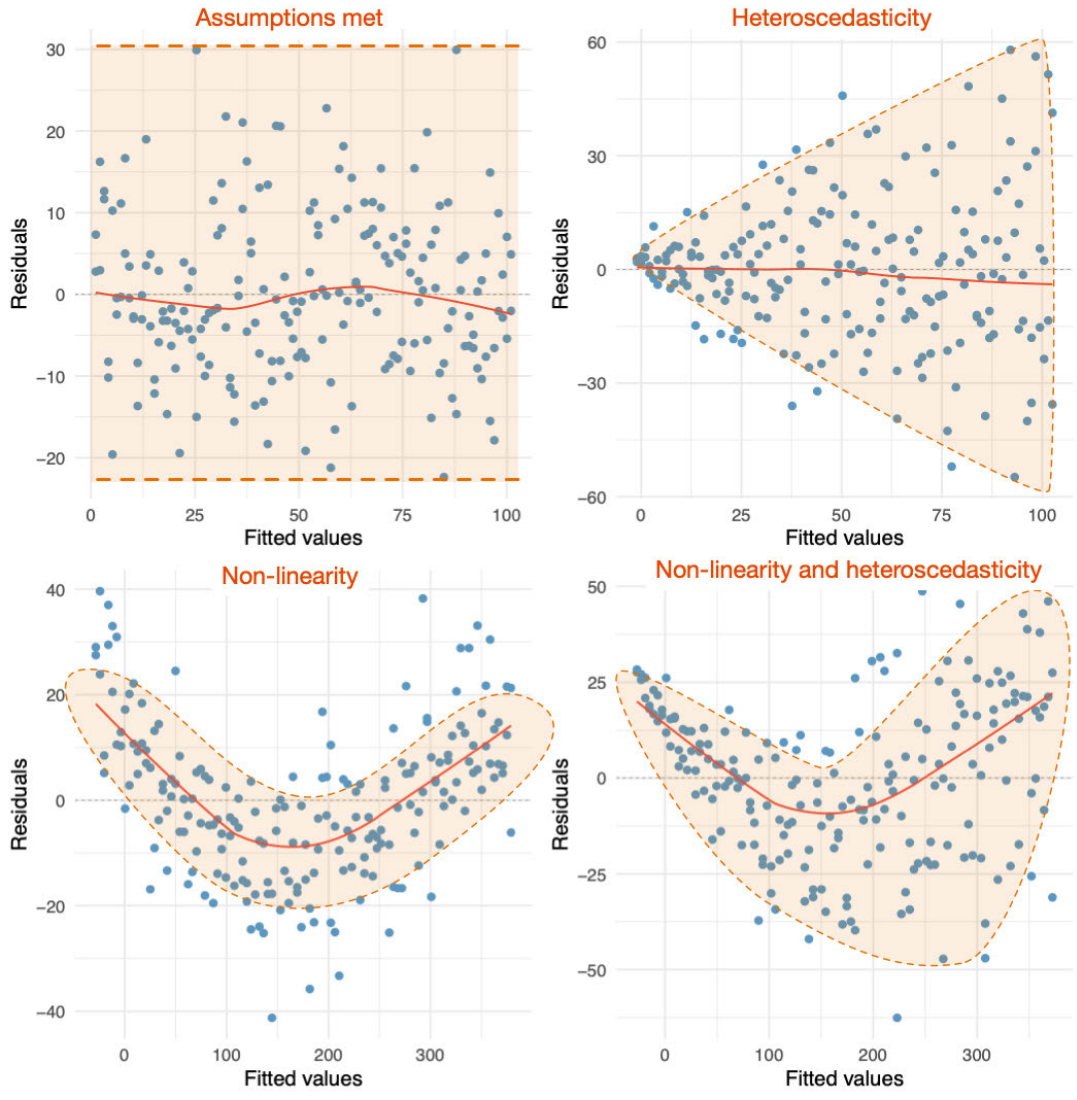
Violation of the assumption

- *b*s are unbiased but not optimal
- Standard error is incorrect
  - Therefore, *t*-tests, *p*-values and confidence intervals will also be incorrect

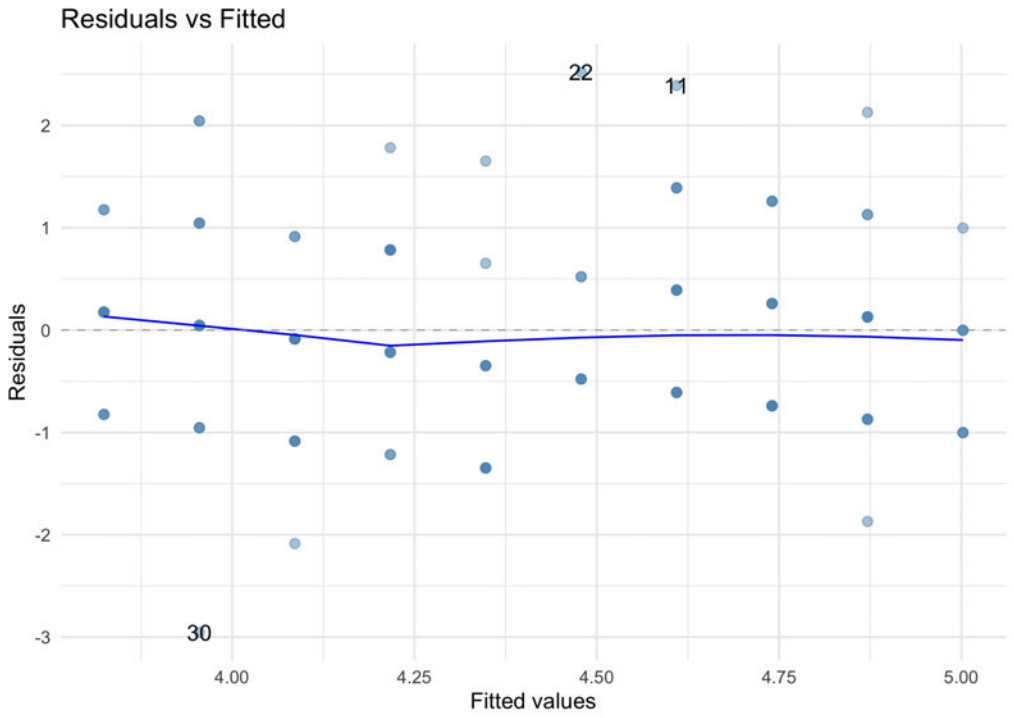
Residuals vs. Predicted values



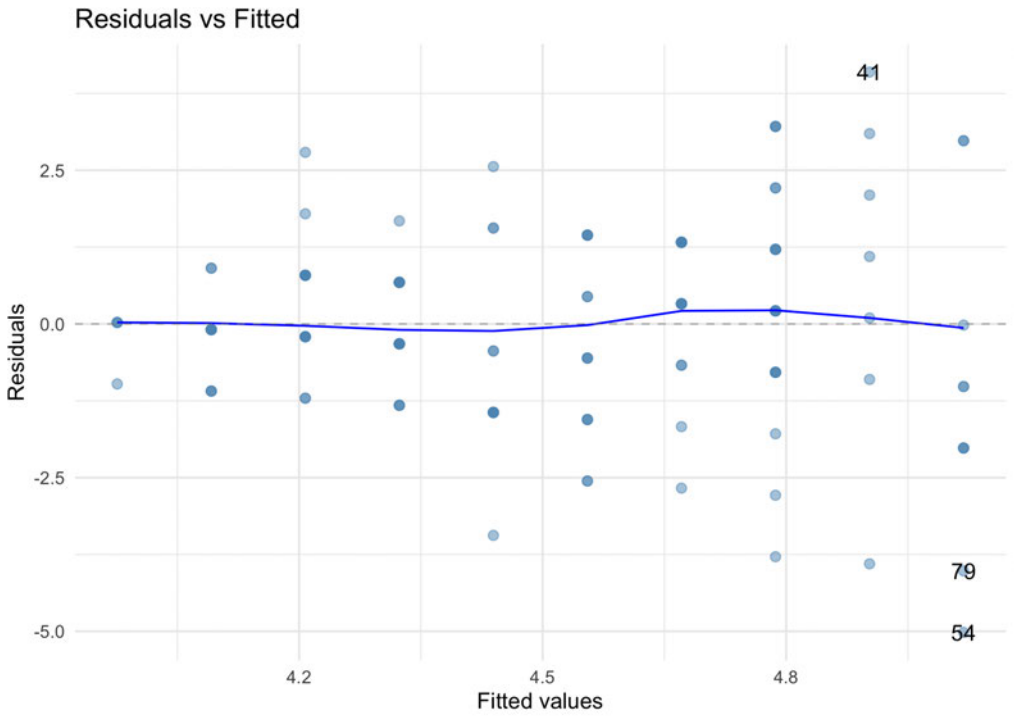
Residuals vs. Predicted values



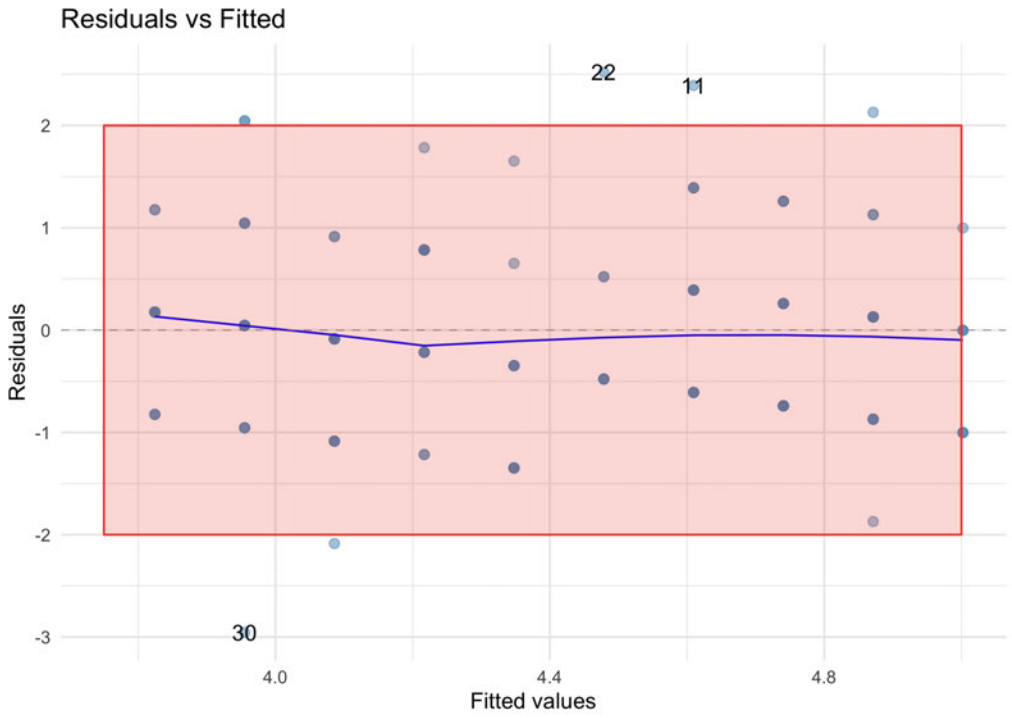
Homoscedastic



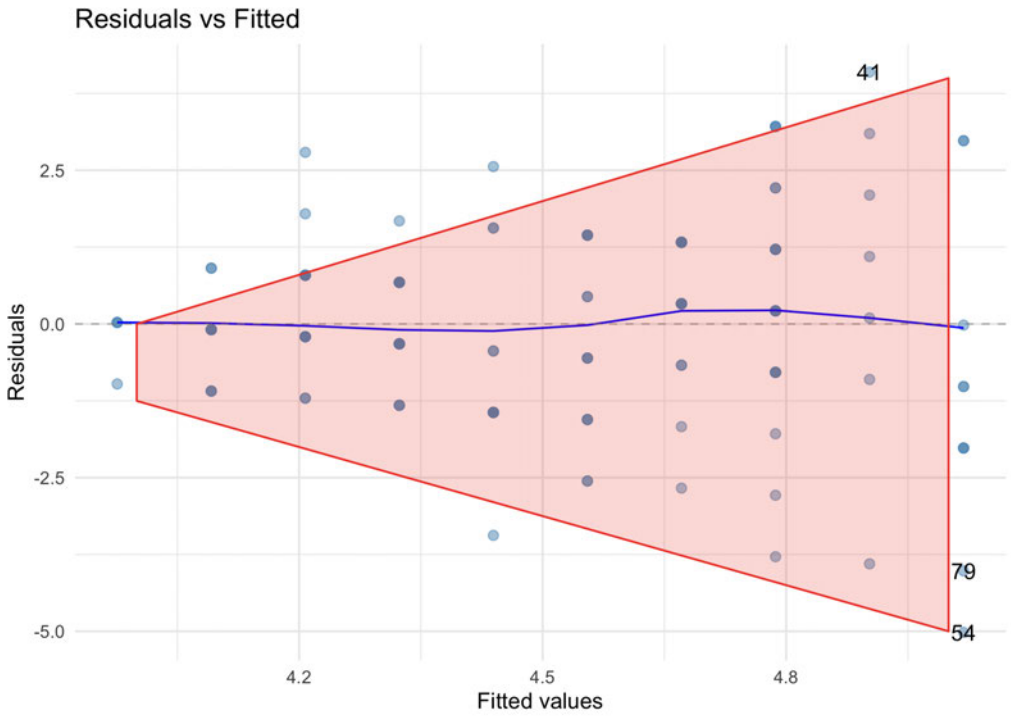
Our data



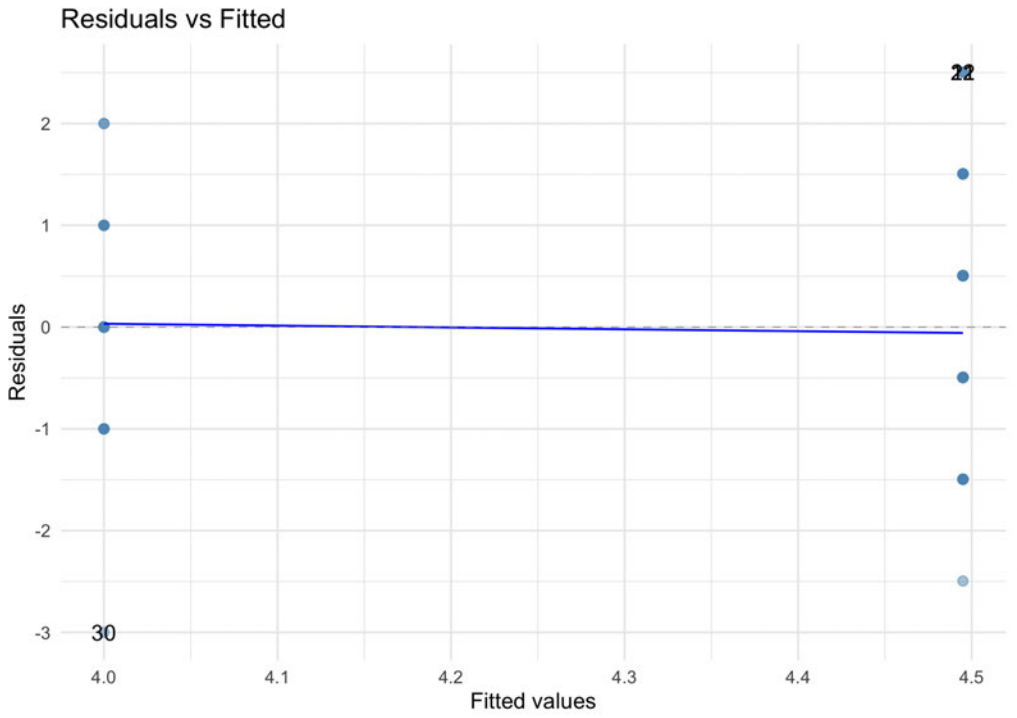
Homoscedastic



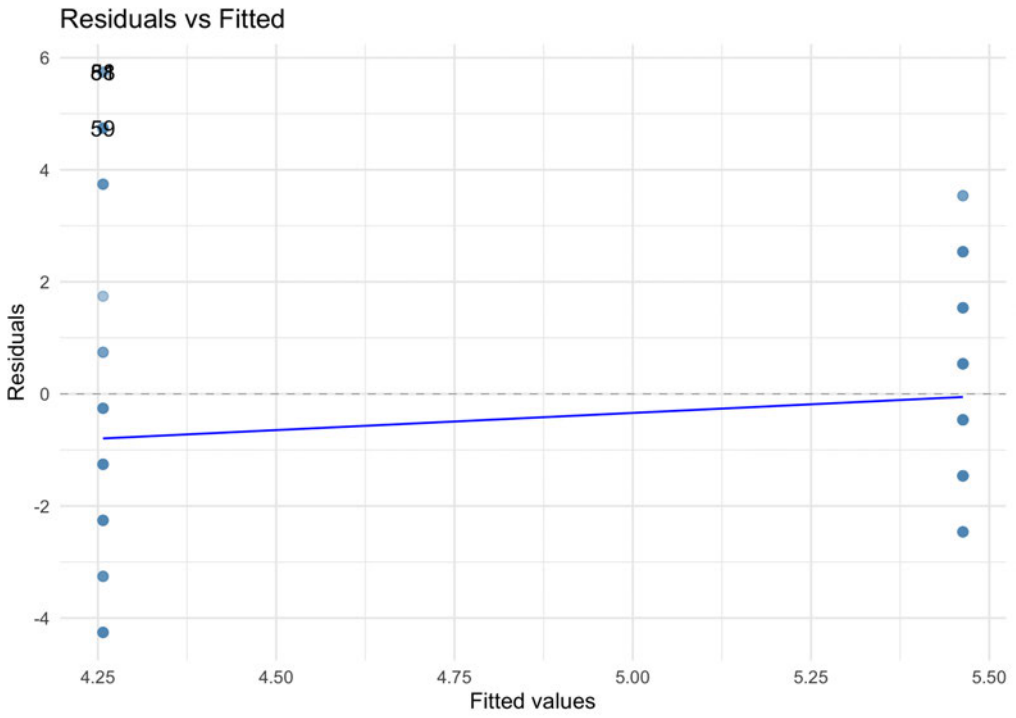
Our data: Heteroscedastic



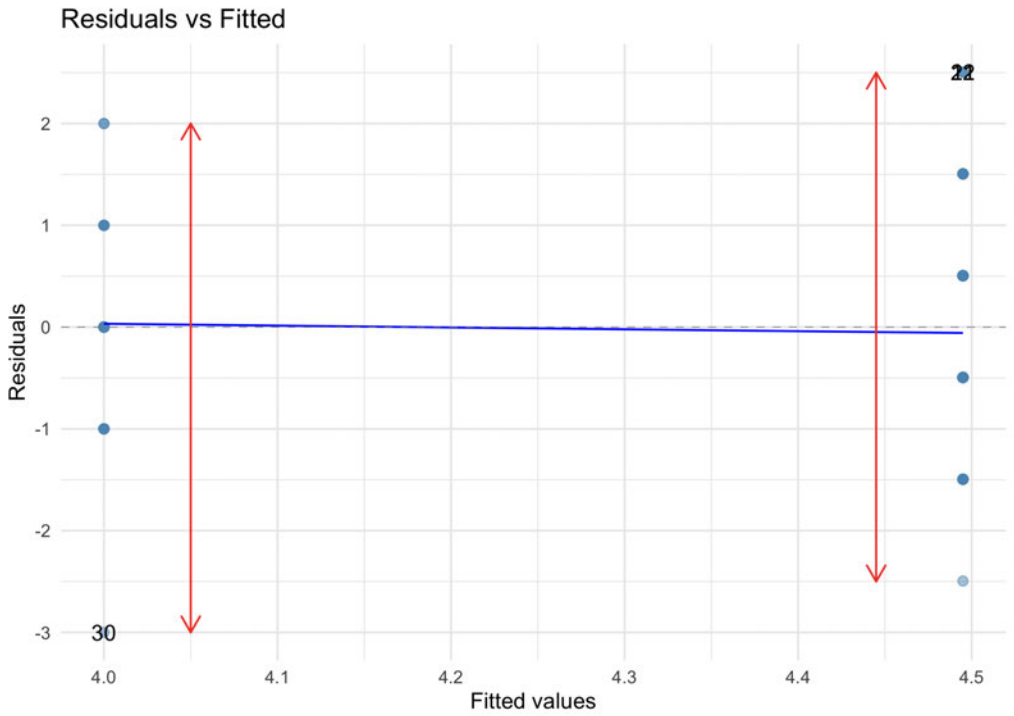
Homoscedastic



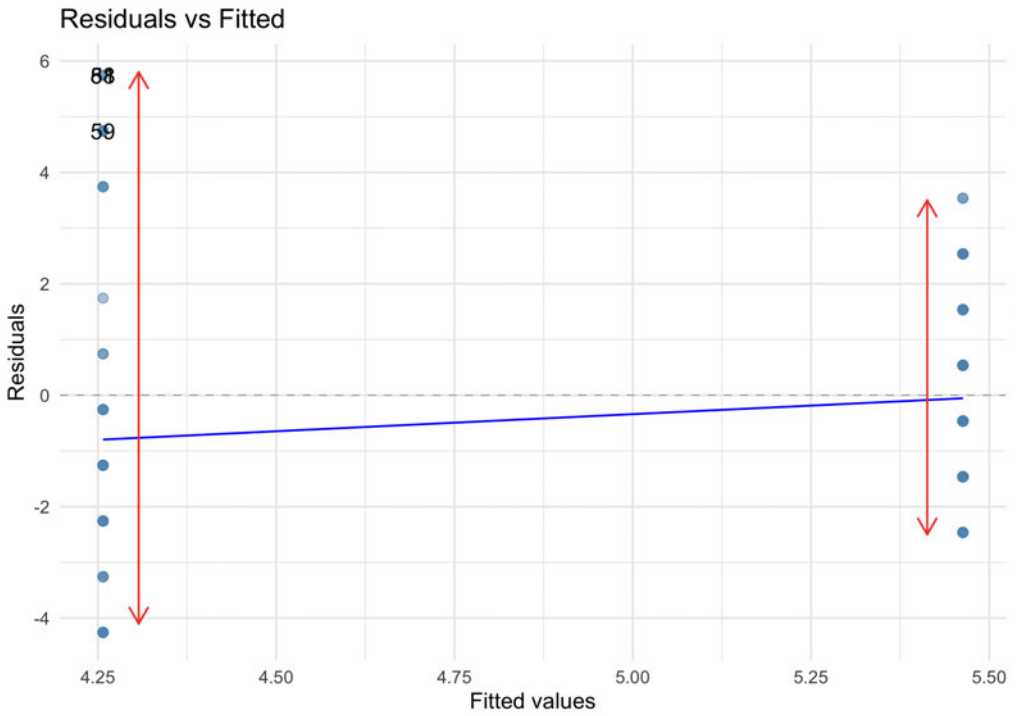
Our data



Homoscedastic



Our data: Heteroscedastic



The Heteroscedasticity Song

Heteroscedasticity is hard to say

If you get it, you'll hope it goes away

Or perhaps that's syphilis, it's hard to tell

But syphilis won't leave you in Statistics Hell

If your residuals are funnelling out

You better get ready to scream and shout

Coz if your data are hetroscedastic

Your model is a lousy fit

Levene's Test

You might hear about it but don't use it: think about what we know about sample sizes and significance.

For the avoidance of doubt, don't use it.

One more time ... don't use it

Part 3: Normality and does dragons poo kill crops?

"Coz their dung doesn't help our crops to grow"

Sir Knight Zach, Defender of the world of Military

Normally distributed something-or-other

People usually (falsely) think the data or population need to be normally distributed

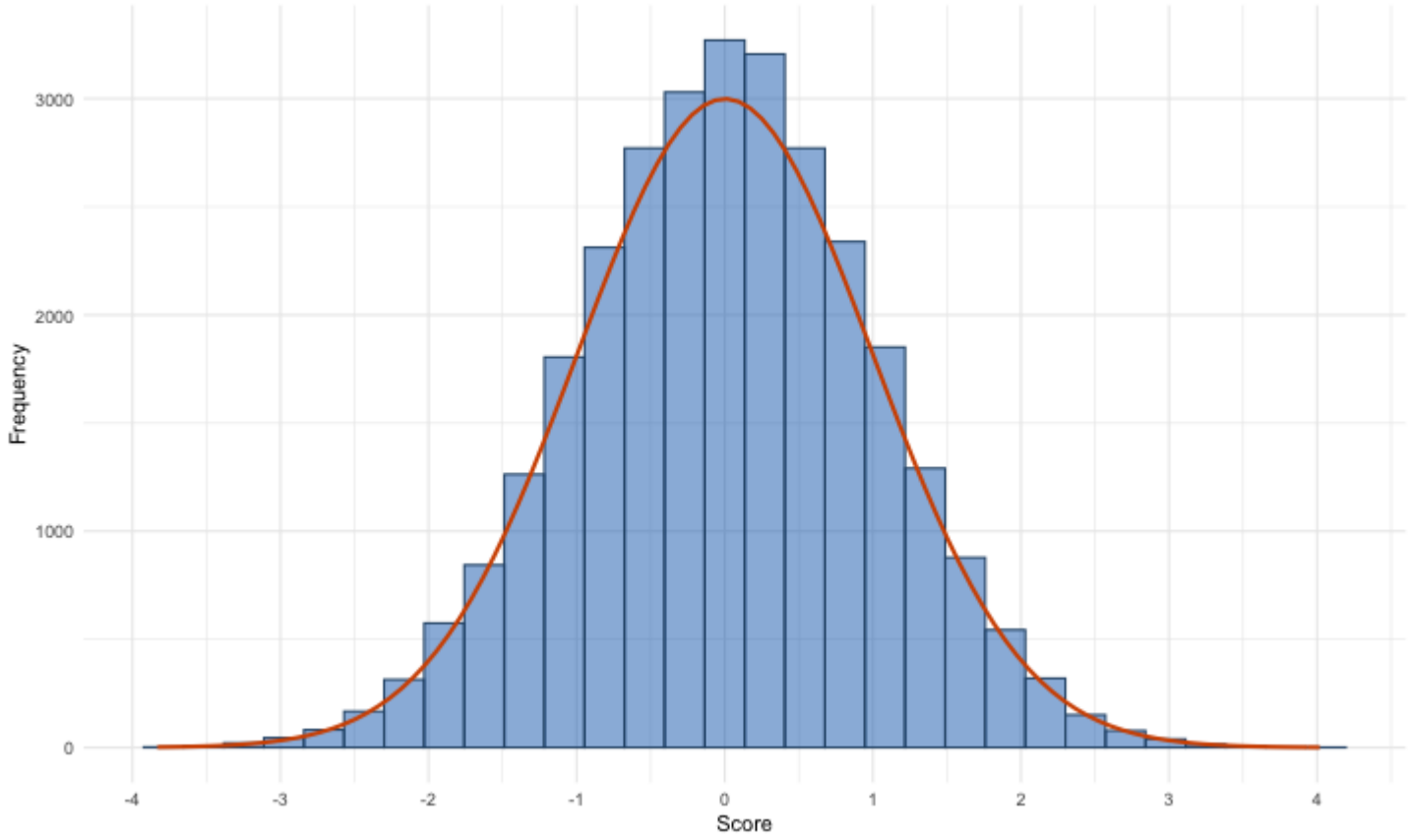
Normality of model errors

- Doesn't really matter
- When errors are not normally distributed, *b* will be unbiased and optimal (i.e., will minimize the variance), but there may be classes of estimator (other than OLS) that are more accurate (Wilcox, 2010)

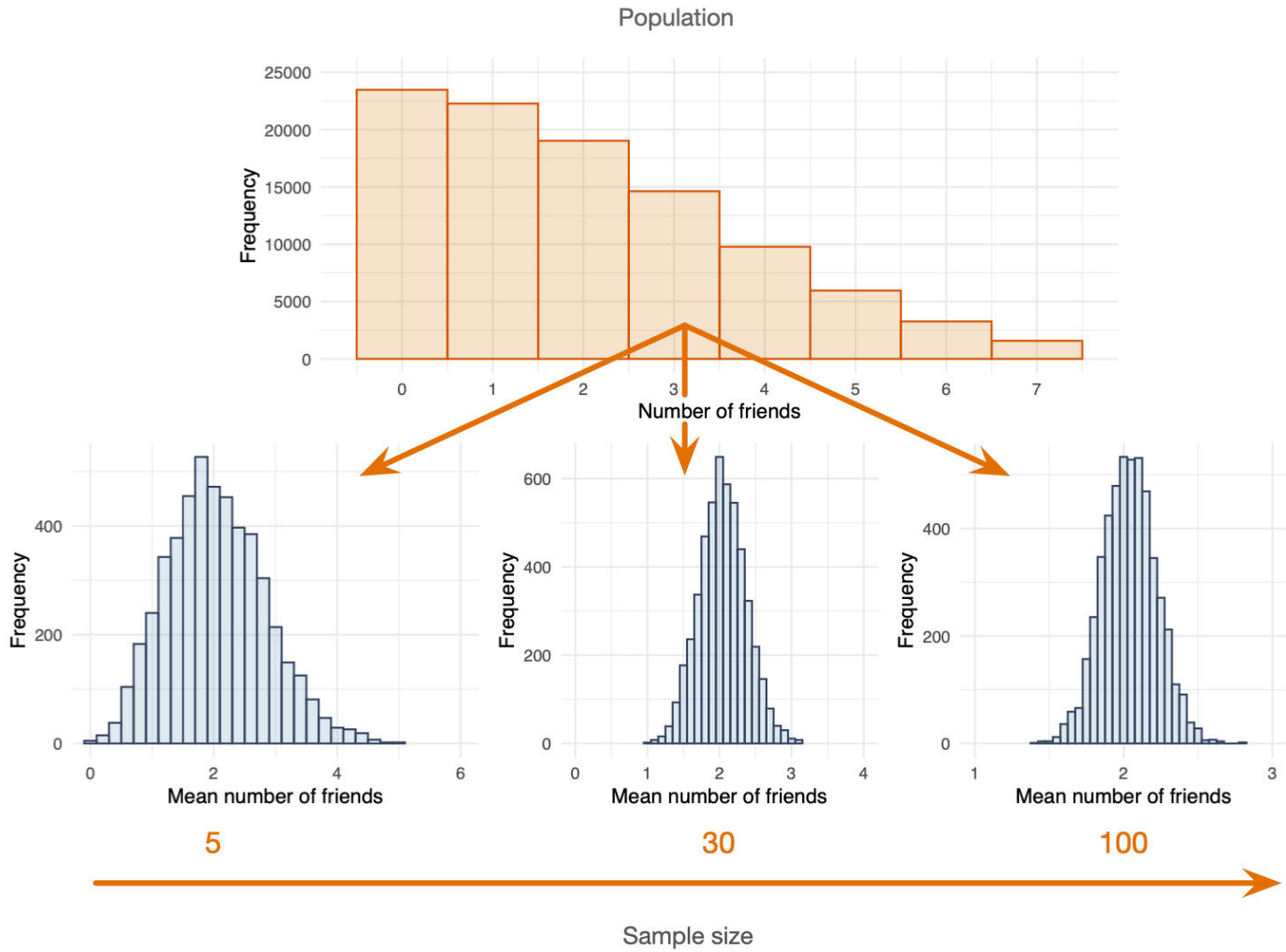
Normality of the sampling distribution

- *p*-values associated with the *b*s of the model assume that the test statistic associated with them follows a normal distribution (or some variant such as *t*)
- Confidence intervals for *b*s are, likewise, constructed using the standard error, which is derived from a sampling distribution of *b*, that is assumed to be normal

A Normal Distribution



The Central Limit Theorem (CLT)



Exploring normality

Model errors

- Check the distribution of the model residuals using a P-P/Q-Q plot

Sampling distribution

- Don't need to worry in large samples because of the CLT
- Use a bootstrap in small samples (more on this later ...)

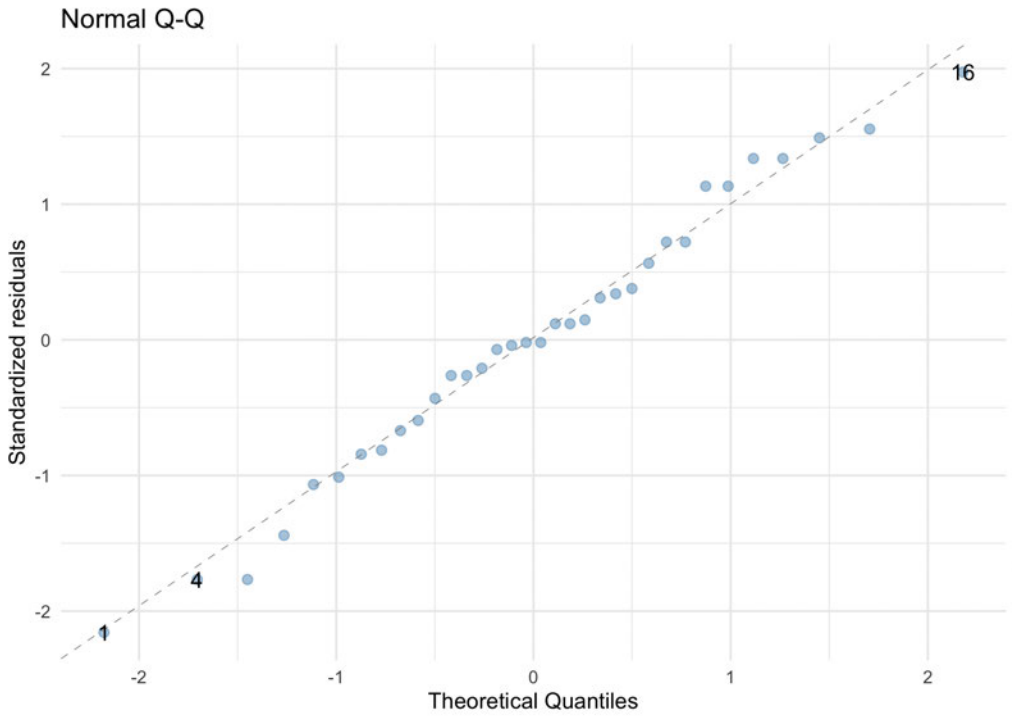
The K-S Test

You might hear about it but don't use it: think about what we know about sample sizes and significance.

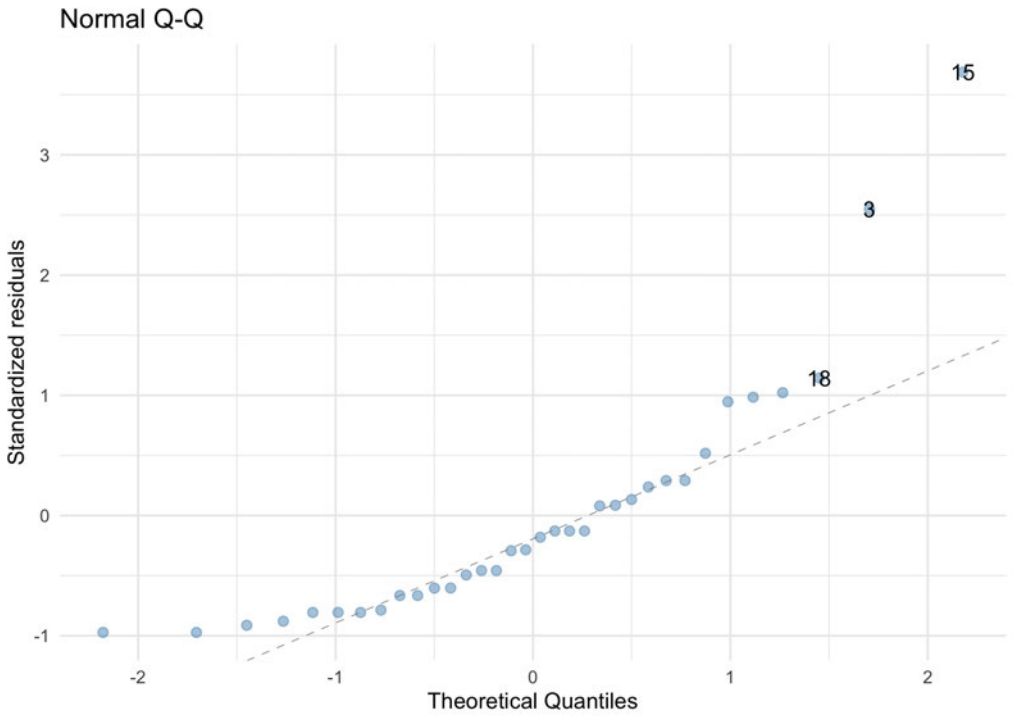
For the avoidance of doubt, don't use it.

One more time ... don't use it

Normal residuals



Our residuals: Non-normal



The Normality Song

If you're feeling the stress 'cos your data's a mess and you're going as mad as a hatter

Coz you really want *p* , to reflect reality. That's when normality matters.

When you've been up for days, in a statistical haze. You're tired and emotionally shattered.

You don't want to be fooled, by your confidence intervals. That's when normality matters

If your life has got skew, and you're wonderin' what to do, 'Coz you feel like your brain has been battered

If your sample is small, then remember the rule: That's when normality matters.

If the scores you collect, are distributional wrecks, remember that this doesn't matter

'Coz for CIs and *p* s, you need normality of the sampling distribution of the parameter

Part 4: Correcting problems?

Robust procedures

The bootstrap

- Standard errors are derived empirically using a resampling technique
- Results in robust confidence intervals and *p*-values
- Designed for small samples (when normality matters)

Heteroskedasticity-consistent standard errors

- Use a sandwich estimator
- HC3 and HC4 methods work best

The Bootstrap

yield	0	6	6	7	7	7	7	7	8	8	12	13	13	15	15	16	16	16	20	22	22	23	26	28	28	29	31	32	40	44	47	48	69	95

Mean = 23.03

Bootstrap sample 1:

yield	0	6	6	7	7	7	7	7	7	7	7	7	8	8	12	13	16	20	22	22	22	23	26	26	28	28	29	31	32	44	47	48	69	69

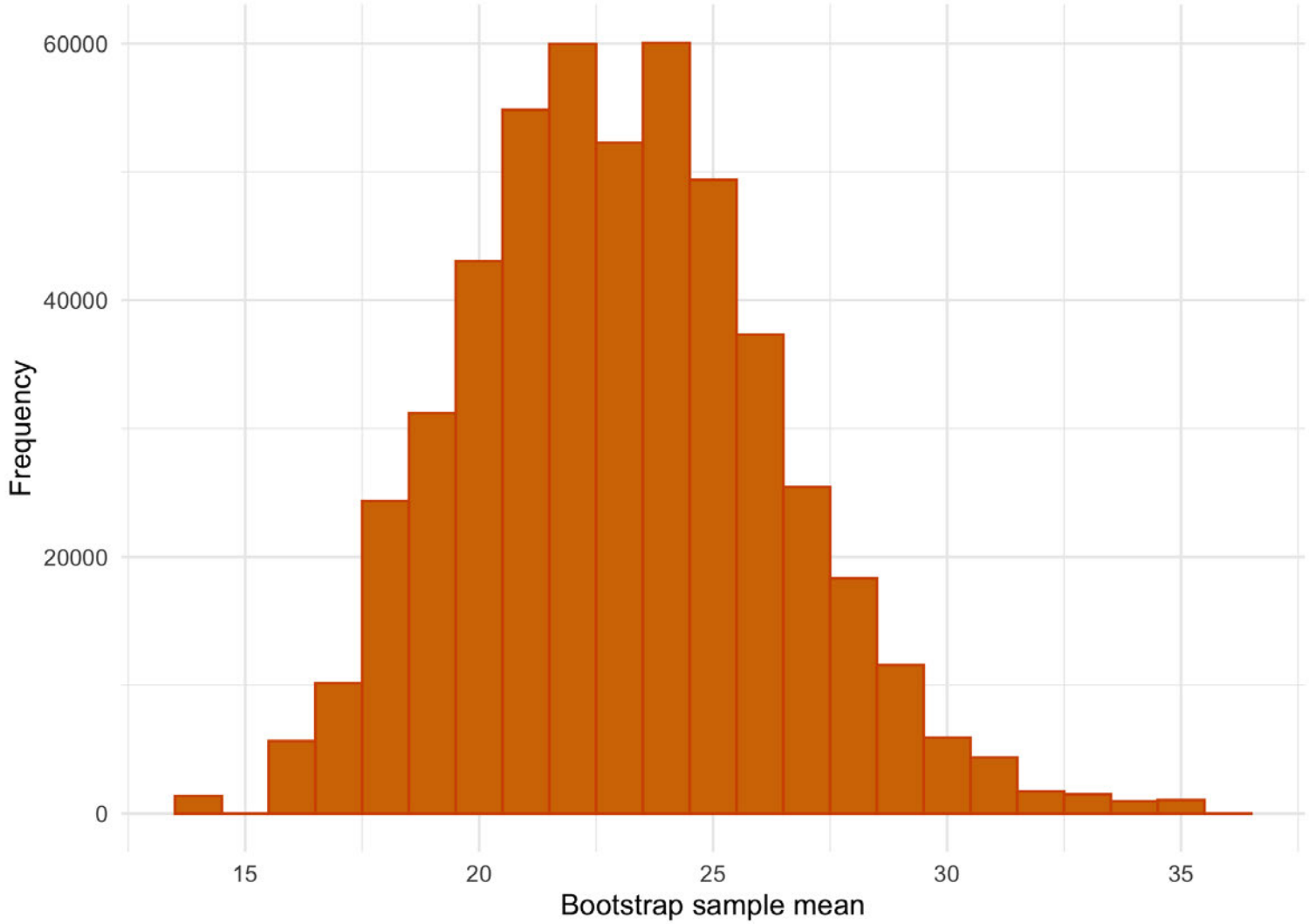
Mean = 21.12

Bootstrap sample 2:

yield	0	0	0	6	6	6	6	7	7	7	8	8	13	13	13	13	15	15	16	16	16	16	20	20	22	23	28	28	29	40	47	48	48	48

Mean = 17.88

yield	0	6	6	7	7	7	7	7	8	8	12	13	13	15	15	16	16	16	20	22	22	23	26	28	28	29	31	32	40	44	47	48	69	95



Do dragons really kidnap royalty?

Normal model (number of dragons)

```
kidnap_lm <- lm(royalty ~ dragons,data = hov_cont_tib)
broom::tidy(kidnap_lm, conf.int = T)
```

term	estimate	std.error	statistic	p.value	conf.low	conf.high
(Intercept)	3.976	0.298	13.323	0.00	3.385	4.566
dragons	0.116	0.056	2.074	0.04	0.005	0.226

Robust model (number of dragons) HC4 Standard errors

```
kidnap_lm <- lm(royalty ~ dragons,data = hov_cont_tib)
parameters::parameters(kidnap_lm, robust = TRUE, vcov.type = "HC4")
```

Parameter	Coefficient	SE	CI_low	CI_high	t	df_error	p
(Intercept)	3.976	0.242	3.497	4.455	16.427	128	0.000
dragons	0.116	0.064	-0.011	0.243	1.805	128	0.073

Normal model (dragons or not)

```
kidnap_gp_lm <- lm(royalty ~ dragons, data = hov_cat_tib)
broom::tidy(kidnap_gp_lm, conf.int = T)
```

term	estimate	std.error	statistic	p.value	conf.low	conf.high
(Intercept)	4.257	0.414	10.281	0.000	3.434	5.080
dragonsDragons	1.206	0.532	2.268	0.026	0.149	2.262

Robust model (dragons or not) HC4 standard errors

```
kidnap_gp_lm <- lm(royalty ~ dragons, data = hov_cat_tib)
parameters::parameters(kidnap_gp_lm, robust = TRUE, vcov.type = "HC4")
```

Parameter	Coefficient	SE	CI_low	CI_high	t	df_error	p
(Intercept)	4.257	0.574	3.117	5.398	7.418	87	0.000
dragonsDragons	1.206	0.616	-0.019	2.431	1.957	87	0.054

Does dragon poo kill crops?

Normal model

```
poop_lm <- lm(yield ~ poop, data = poop_tib)
broom::tidy(poop_lm, conf.int = T)
```

term	estimate	std.error	statistic	p.value	conf.low	conf.high
(Intercept)	14.343	6.765	2.120	0.042	0.563	28.124
poop	2.023	1.371	1.476	0.150	-0.770	4.815

Bootstrap model

```
poop_lm <- lm(yield ~ poop, data = poop_tib)
parameters::parameters(poop_lm, bootstrap = TRUE)
```

Parameter	Coefficient	CI_low	CI_high	p
(Intercept)	14.144	5.117	25.920	0.002
poop	2.058	0.086	3.919	0.046

Summary

The key assumptions of the General Linear Model and what they affect are (in order of importance)

- Linearity and additivity
  - If you don't have these then you're fitting the wrong model in the first place
- Spherical errors (homoscedastic and independent errors). When violated:
  - *b* s are unbiased but not optimal
  - Standard error of parameter, associated *t*-test, *p*-value and confidence intervals will be incorrect
- Normality of residuals and sampling distribution
  - Normality of errors doesn't *really* matter
  - Normality of sampling distribution matters for *p*-values and confidence intervals associated with the *b*s of the model.
  - Central limit theorem!
- Pay attention to outliers and influential cases
  - Standardized residuals
  - Cook's distance (absolute value > 1)
- Robust methods
  - Bootstrapping
  - Use robust SEs