Modelling fundamentals: a linear model with a single, continuous X

BIOL2022 - L02b

Dr Januar Harianto The University of Sydney

THE UNIVERSITY OF SYDNEY

Learning objectives

By the end of this lecture, you should be able to:

- [ ] Describe a linear model for a continuous response and predictor.
- [ ] Write and interpret a simple linear model equation.
- [ ] Interpret coefficients, confidence intervals, and R^2 in model output.

Introduction to linear modelling

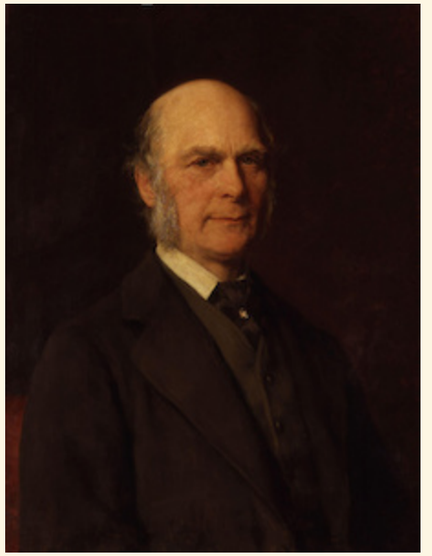
> “Cars with flames painted on the hood might get more speeding tickets. Are the flames making the car go fast? No. Certain things just go together. And when they do, they are correlated. **It is the darling of all human errors to assume, without proper testing, that one is the cause of the other.**”

— Barbara Kingsolver, Flight Behavior (2012)

A short history

Illustrative image of Adrien-Marie Legendre

Carl Friedrich Gauss



Francis Galton

Legendre: Introduced least squares in 1806. Best image that I could find. Source: Wikipedia

Gauss: Published work on least squares in 1809. Source: Wikipedia

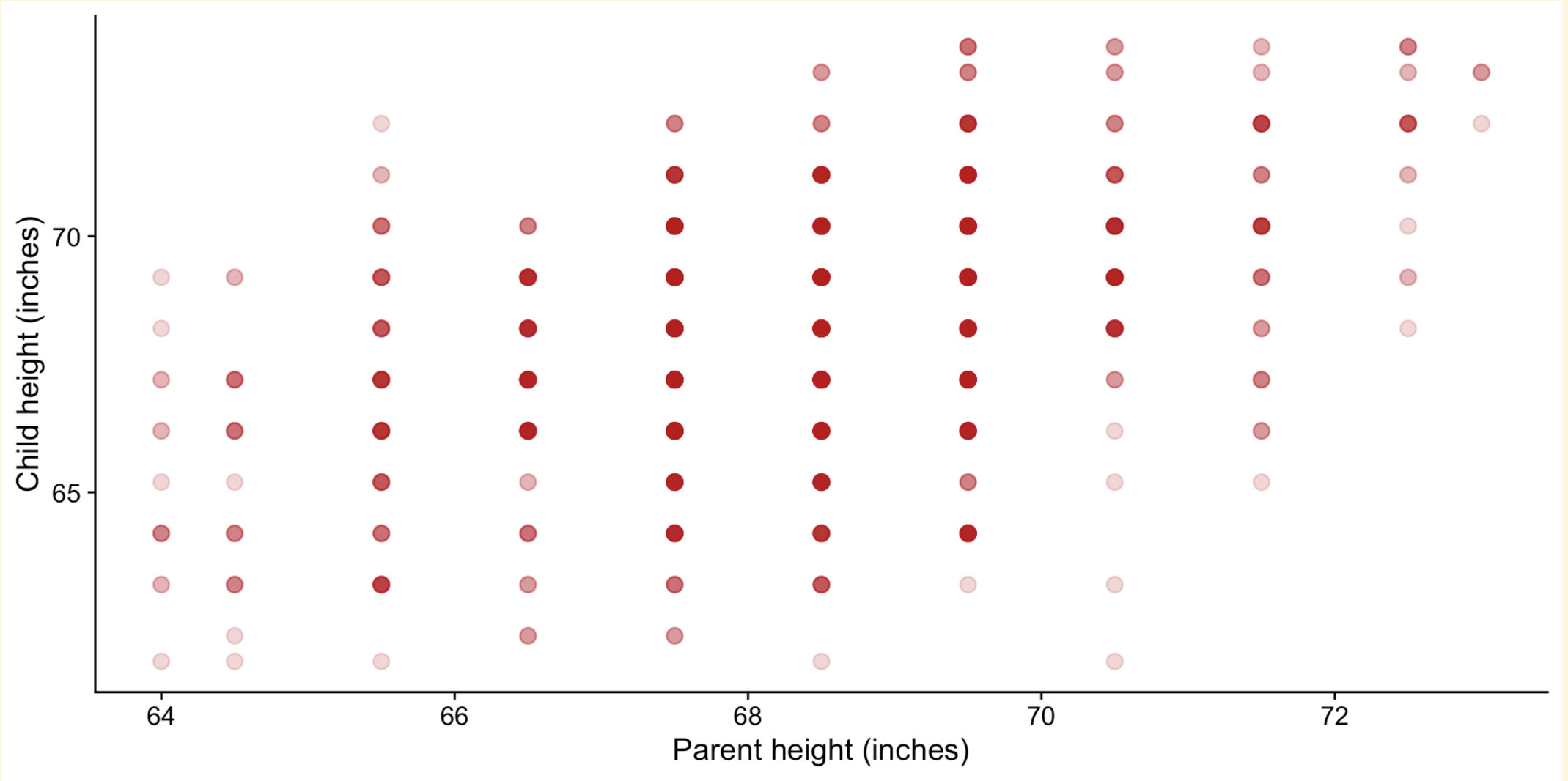
Galton: Introduced regression to the mean (1886) and correlation (1888).

The Galton dataset (as a quick example)

- 928 children from 205 parent pairs.
- Parent and child heights were recorded in inches.
- Heights were recorded in size classes, so the plot will show horizontal bands.

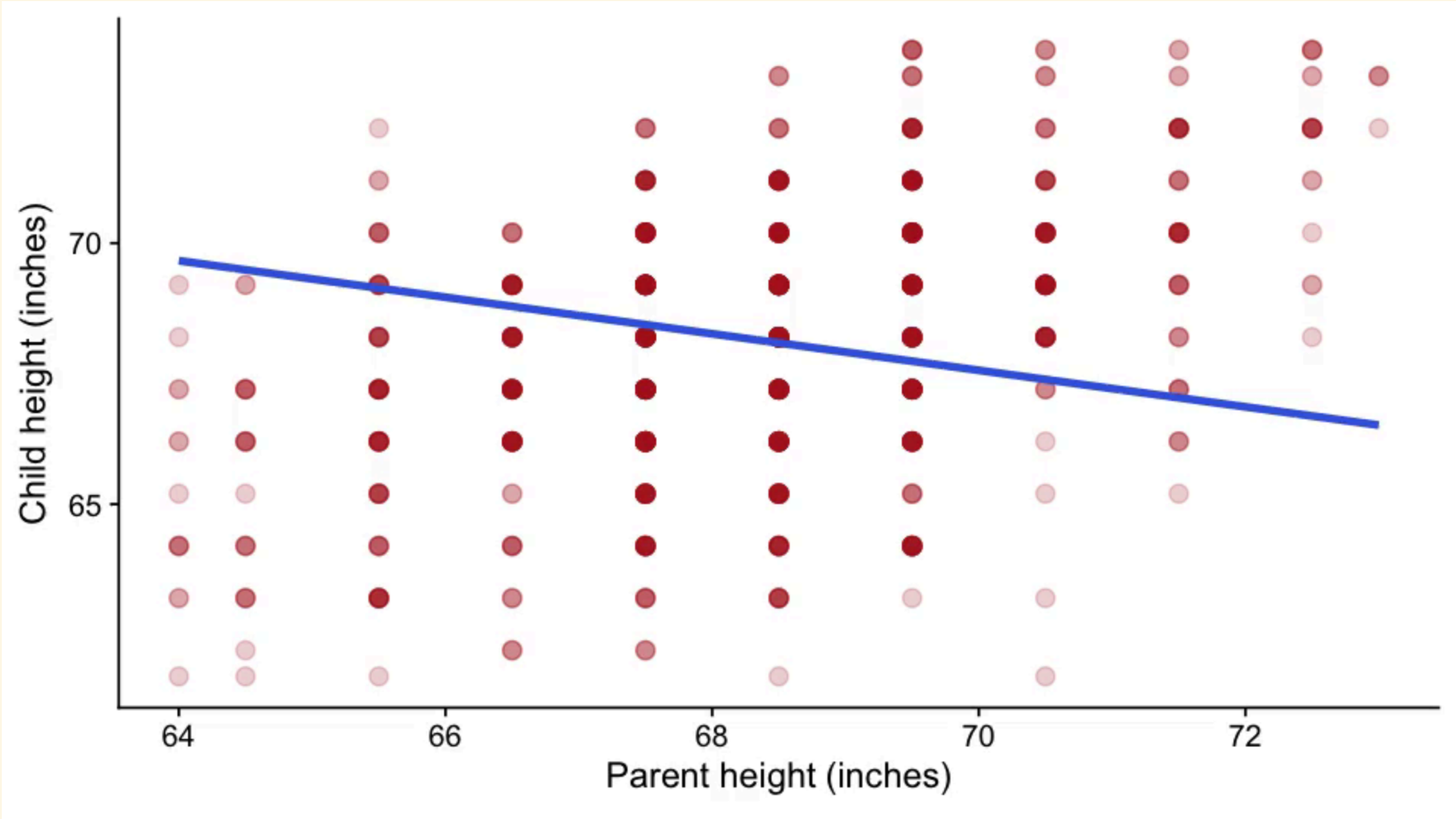
Question: How does child height vary with parent height?

Scatterplot



We want to explain this *noisy* relationship…

Regression line



We want to explain this *noisy* relationship… **using a linear model…?**

How does linear modelling work?

Least squares

> The method of least squares is the **automobile of modern statistical analysis**: despite its limitations, occasional accidents and incidental pollution, it and its numerous variations, extensions, and related conveyances **carry the bulk of statistical analyses**, and are known and valued by nearly all.

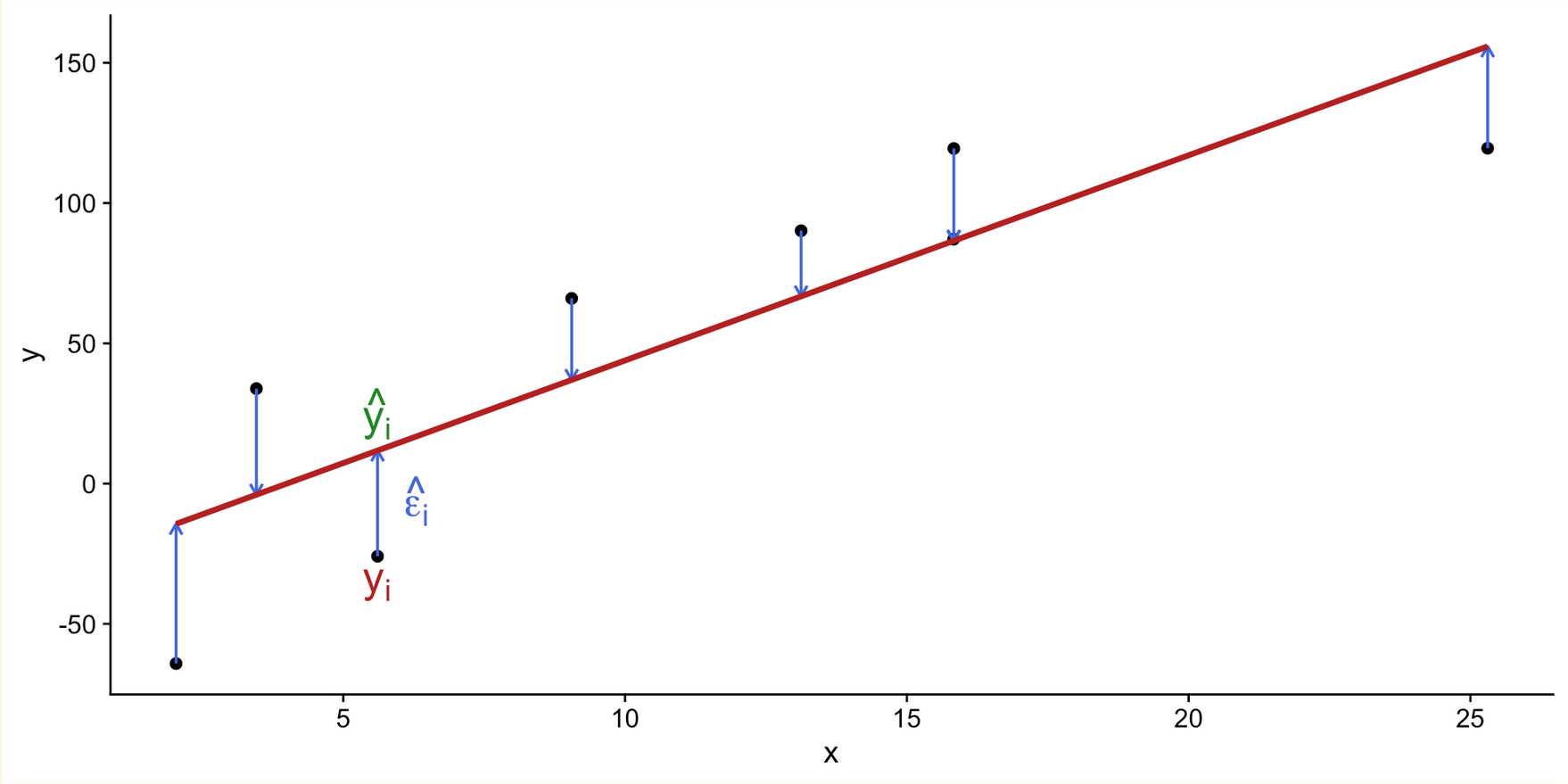
– Stigler, 1981 (emphasis added)

Residuals, $\hat{\epsilon}$

$$Residual = observed - predicted$$

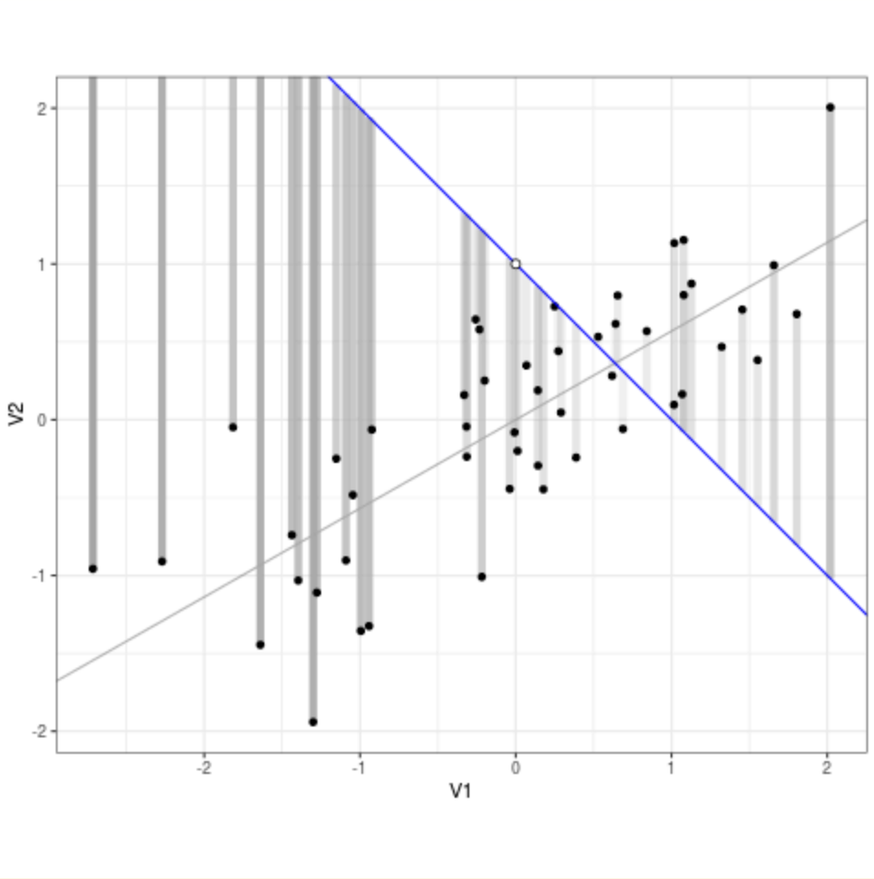
$$\hat{\epsilon}_i = y_i - \hat{y}_i$$

If a child is 178 cm tall and the fitted line predicts 174 cm, the residual is +4 cm.


150
100
50
0
-50
y
5
10
15
20
25
x
ŷi
ε̂i
yi

Using residuals to find the best line

Choose the intercept (β_0) and slope (β_1) that make the total squared residual as small as possible:



$$\underset{\beta_0, \beta_1}{\arg\min} \sum_{i=1}^{n} (y_i - (\beta_0 + \beta_1 x_i))^2$$

1. Draw a candidate line.
2. Calculate the residuals.
3. Square and sum them.
4. Find the line with the smallest sum.
5. **Report its slope and intercept.**

Source: ls-springs

The linear regression

Model definition

A linear model describes the relationship between a response y and a predictor x.

$$y_i = \beta_0 + \beta_1 x_i + \epsilon_i$$

- β_0: intercept
- β_1: slope
- ϵ_i: deviation of observation i from the mean response predicted at x_i

Understanding the model

In the model:

$$y_i = \beta_0 + \beta_1 x_i + \epsilon_i$$

The blue component is the fixed mean relationship. The red component is random variation around it.

We can think of the response as being made up of two components:

- Response = Prediction + Error
- Response = Signal + Noise
- Response = Deterministic + Random
- Response = Explainable + Everything else

How do we "explain" a fitted model?

$$y_i = \beta_0 + \beta_1 x_i + \epsilon_i$$

If we were to use plain English, we could say:

> The response variable y changes by β_1 units for every 1-unit change in the predictor x, and the average value of y when $x = 0$ is β_0. The remaining variation in y is due to random error.

From model to fitted equation

The model:

$$y_i = \beta_0 + \beta_1 x_i + \epsilon_i$$

After fitting the data:

$$\hat{y}_i = \hat{\beta}_0 + \hat{\beta}_1 x_i$$

For Galton's data:

$$\widehat{child} = 23.94 + 0.65(parent)$$

Once we fit the model, it is now called an estimate (although other terms are also used, such as "fitted value" or "predicted value"), and we no longer have the error term because we are now describing the mean response at each value of x.

Interpretation?

$$\widehat{child} = 23.94 + 0.65(parent)$$

Term	Estimate (SE)	95% CI	p-value
Coefficients			
Intercept	23.94 (2.81)	18.43–29.46	<0.001
Parent height (in)	0.65 (0.041)	0.57–0.73	<0.001
Model fit			
R^2	0.21		
Adjusted R^2	0.21		
n	928		

SE = standard error; CI = confidence interval.

$$\widehat{child} = 23.94 + 0.65(parent)$$

Term	Estimate (SE)	95% CI	p-value
Coefficients			
Intercept	23.94 (2.81)	18.43–29.46	<0.001
Parent height (in)	0.65 (0.041)	0.57–0.73	<0.001
Model fit			
R^2	0.21		
Adjusted R^2	0.21		
n	928		

SE = standard error; CI = confidence interval.

A 1-inch increase in parent height corresponds to a 0.65-inch increase in predicted mean child height. This value corresponds to $\hat{\beta}_1$, the estimated slope of the fitted line.

$$\widehat{child} = 23.94 + 0.65(parent)$$

Term	Estimate (SE)	95% CI	p-value
Coefficients			
Intercept	23.94 (2.81)	18.43–29.46	<0.001
Parent height (in)	0.65 (0.041)	0.57–0.73	<0.001
Model fit			
R^2	0.21		
Adjusted R^2	0.21		
n	928		

SE = standard error; CI = confidence interval.

- If you see SE, that's the standard error of the estimated slope. It describes how precise the estimate is.
- If you see a 95% CI, that indicates the range of slope values we are confident that if we repeated the experiment many times, 95% of the time the true slope would fall within that range.
- The p-value tests the null hypothesis that the slope is zero. A small p-value indicates that the slope is unlikely to be zero, and we can conclude that there is a **significant** linear association between parent and child height.

$$\widehat{child} = 23.94 + 0.65(parent)$$

Term	Estimate (SE)	95% CI	p-value
Coefficients			
Intercept	23.94 (2.81)	18.43–29.46	<0.001
Parent height (in)	0.65 (0.041)	0.57–0.73	<0.001
Model fit			
R^2	0.21		
Adjusted R^2	0.21		
n	928		

SE = standard error; CI = confidence interval.

The intercept is $\hat{\beta}_0$, the estimated mean child height when parent height is zero. Does this make sense?

$$\widehat{child} = 23.94 + 0.65(parent)$$

Term	Estimate (SE)	95% CI	p-value
Coefficients			
Intercept	23.94 (2.81)	18.43–29.46	<0.001
Parent height (in)	0.65 (0.041)	0.57–0.73	<0.001
Model fit			
R^2	0.21		
Adjusted R^2	0.21		
n	928		

SE = standard error; CI = confidence interval.

$R^2 = 0.21$ explains that parent height accounts for 21% of the observed variation in child height. Adjusted R^2 applies a penalty for model complexity. With one predictor and 928 observations, it remains 0.21. Finally, n is the number of observations used to fit the model.

Putting the result into words

$$\widehat{child} = 23.94 + 0.65(parent)$$

1. Statistical interpretaion
2. Biological interpretation

> There was a significant, positive linear association between parent and child height (95% CI 0.57–0.73, $p < 0.001$), and the model accounted for 21% of the variation in child height. A 1-inch increase in parent height corresponded to a 0.65-inch increase in predicted mean child height.

Practice

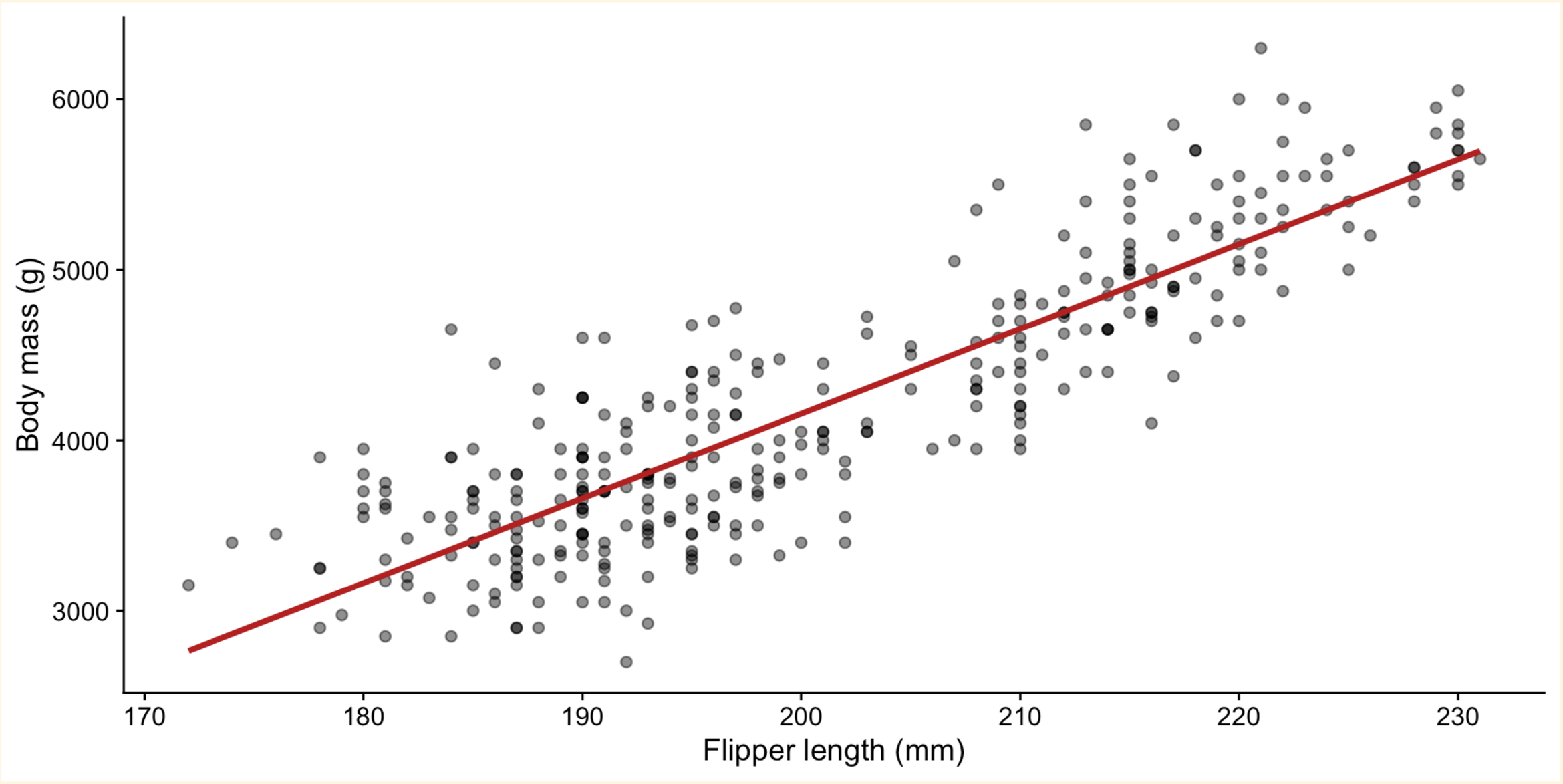
Example: Adelie penguins

```
lm(formula = body_mass_g ~ flipper_length_mm, data = penguins)

Residuals:
    Min      1Q  Median      3Q     Max
-875.68 -331.10  -14.53  265.74 1144.81

Coefficients:
                   Estimate Std. Error t value Pr(>|t|)
(Intercept)       -2535.837    964.798  -2.628  0.00948 **
flipper_length_mm    32.832      5.076   6.468 1.34e-09 ***
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 406.6 on 149 degrees of freedom
  (1 observation deleted due to missingness)
Multiple R-squared:  0.2192,    Adjusted R-squared:  0.214
F-statistic: 41.83 on 1 and 149 DF,  p-value: 1.343e-09
```


Body mass (g)
6000
5000
4000
3000
170
180
190
200
210
220
230
Flipper length (mm)

Example: *Iris setosa*

Linear Regression

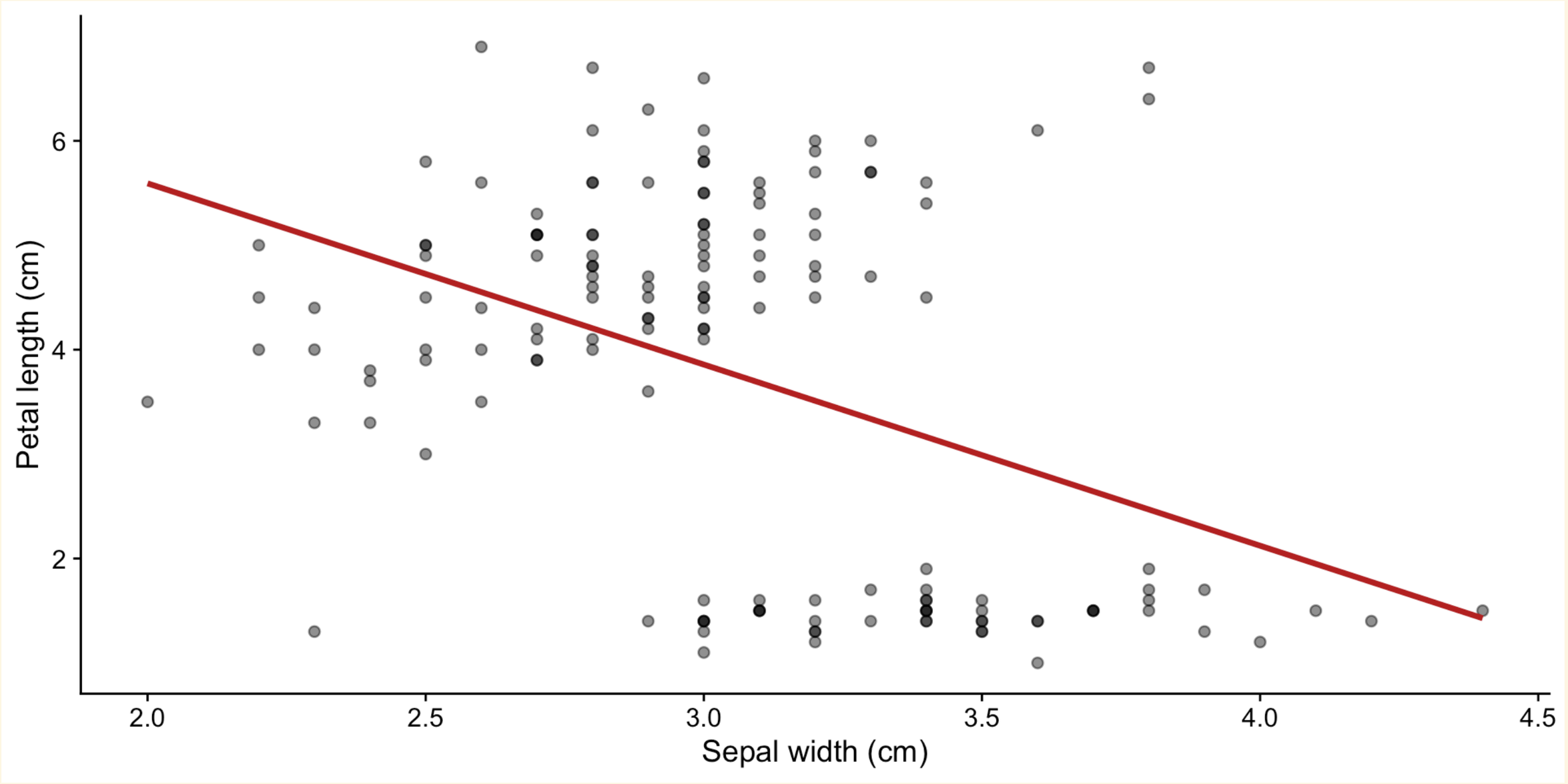
Model Fit Measures

Model	R	R^2
1	0.178	0.0316

Note. Models estimated using sample size of N=50

Model Coefficients - Petal.Length

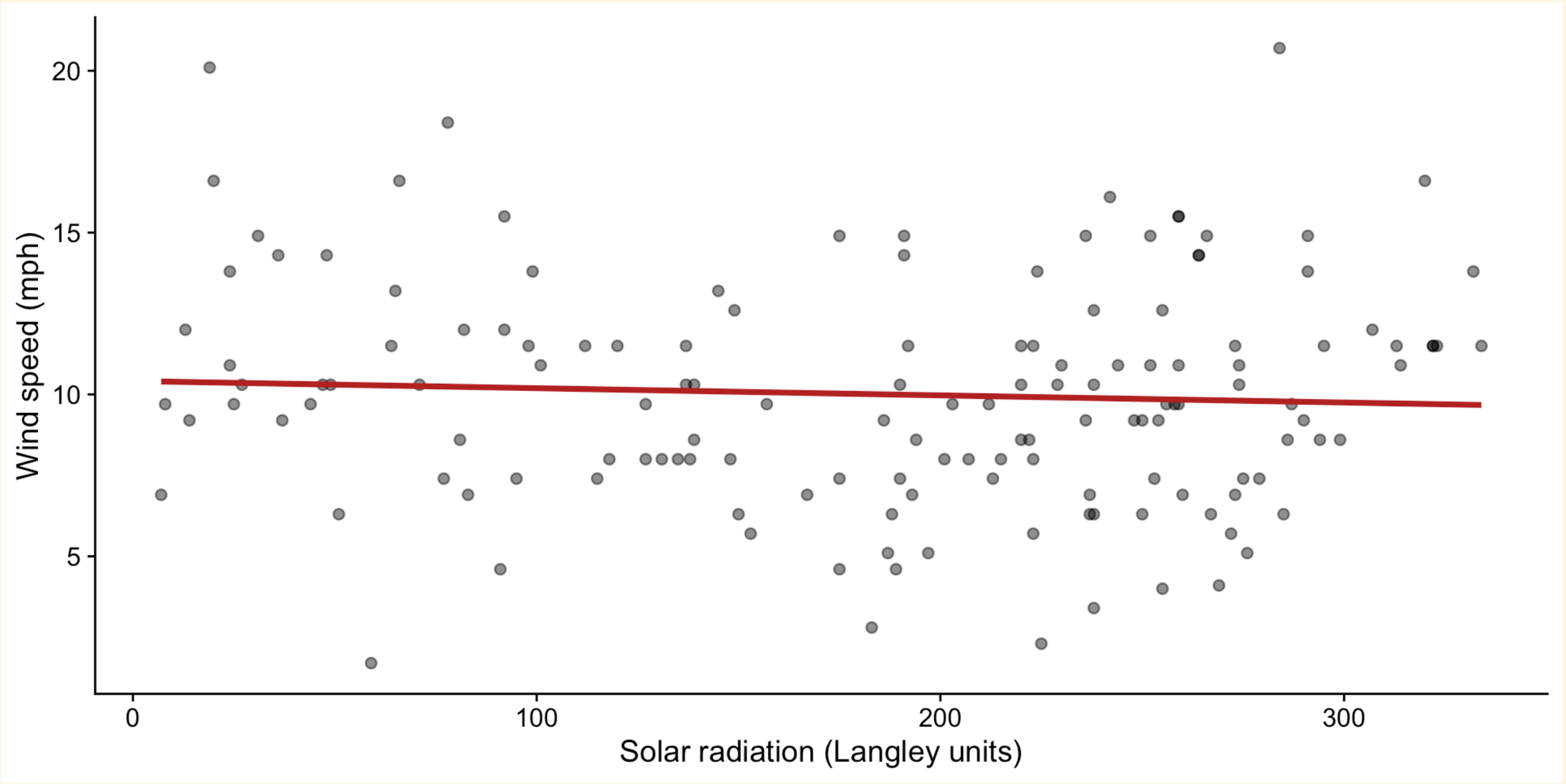
			95% Confidence Interval			
Predictor	Estimate	SE	Lower	Upper	t	p
Intercept	1.1829	0.2244	0.7317	1.634	5.27	<.001
Sepal.Width	0.0814	0.0651	-0.0494	0.212	1.25	.217


Petal length (cm)
6
4
2
2.0
2.5
3.0
3.5
4.0
4.5
Sepal width (cm)

Example: Air quality data

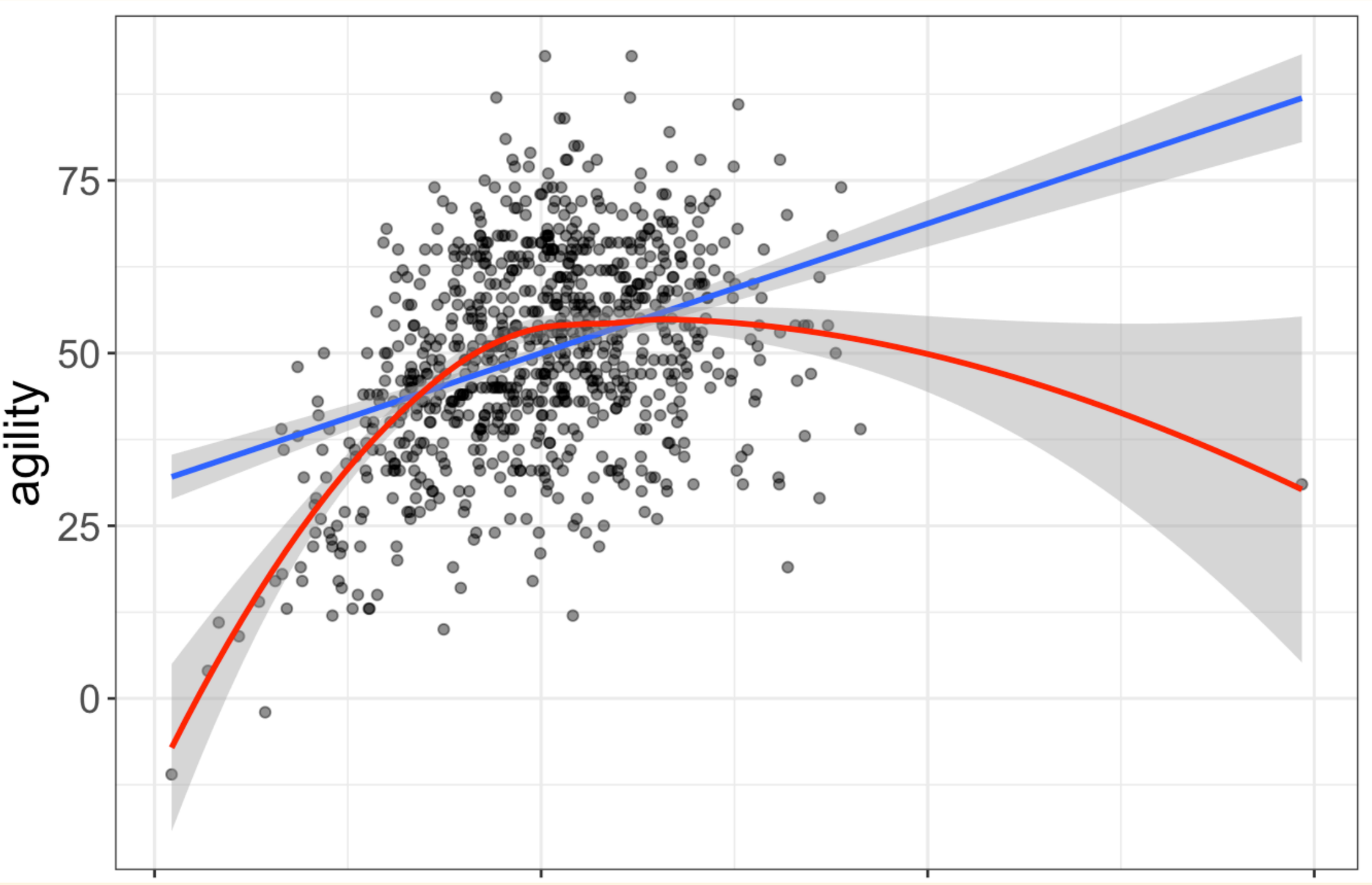
▶ Code

Term	Estimate (SE)	95% CI	p-value
Intercept	10.415 (0.669)	9.092 to 11.738	<0.001
Solar radiation (Langley units)	-0.002 (0.003)	-0.009 to 0.004	0.5
Model fit: R^2 = 0.003; adjusted R^2 = -0.004; n = 146.			


20
15
10
5
Wind speed (mph)
0
100
200
300
Solar radiation (Langley units)

A fitted model is not automatically a good model

Any dataset can be given a fitted straight line, but that does not mean the line describes the data well.


agility
75
50
25
0

Next week: model assumptions

We will explore the assumptions of linear regression and learn how to assess whether they are reasonable for our data.

Thanks