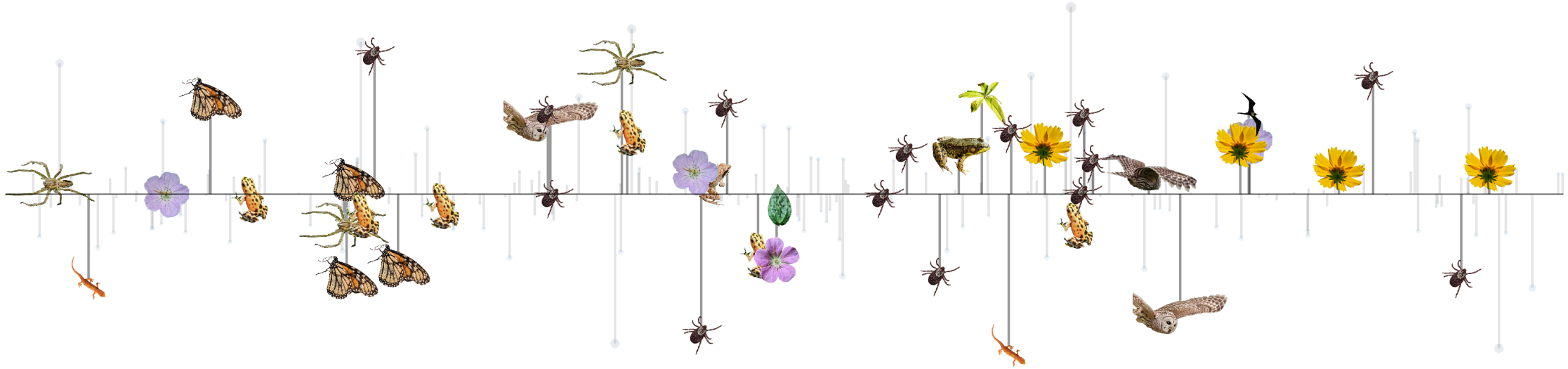
Intro to Quantitative Ecology
UMass Amherst – Michael France Nelson

Deck 9 – Linear Regression 1



Linear Regression Overview

Important Concepts

Regression Overview + Key Concepts

Ordinary Least Squares (OLS)

- OLS Linear regression is the framework for almost all the analyses we have covered so far.
- Even t-tests are part of this world.
- Rich and complex topic.
- Not the only modeling paradigm!

Key themes:

- Model Thinking: broad-picture
- Specific OLS Methods: ANOVA, simple linear regression, etc.
- Model assumptions and validation
- Model interpretation and communication

Linear Regression Concepts

The first batch of OLS (Ordinal Least Squares) concepts we'll cover is:

- Simple Linear Regression
- Multiple Linear Regression
- Model coefficients
- Interpreting model summaries
- Implementing simple- and multiple-linear regression in R
- Linear regression connections to ANOVA (next week)
  - ANOVA is just a linear model in disguise

Correlation and Regression

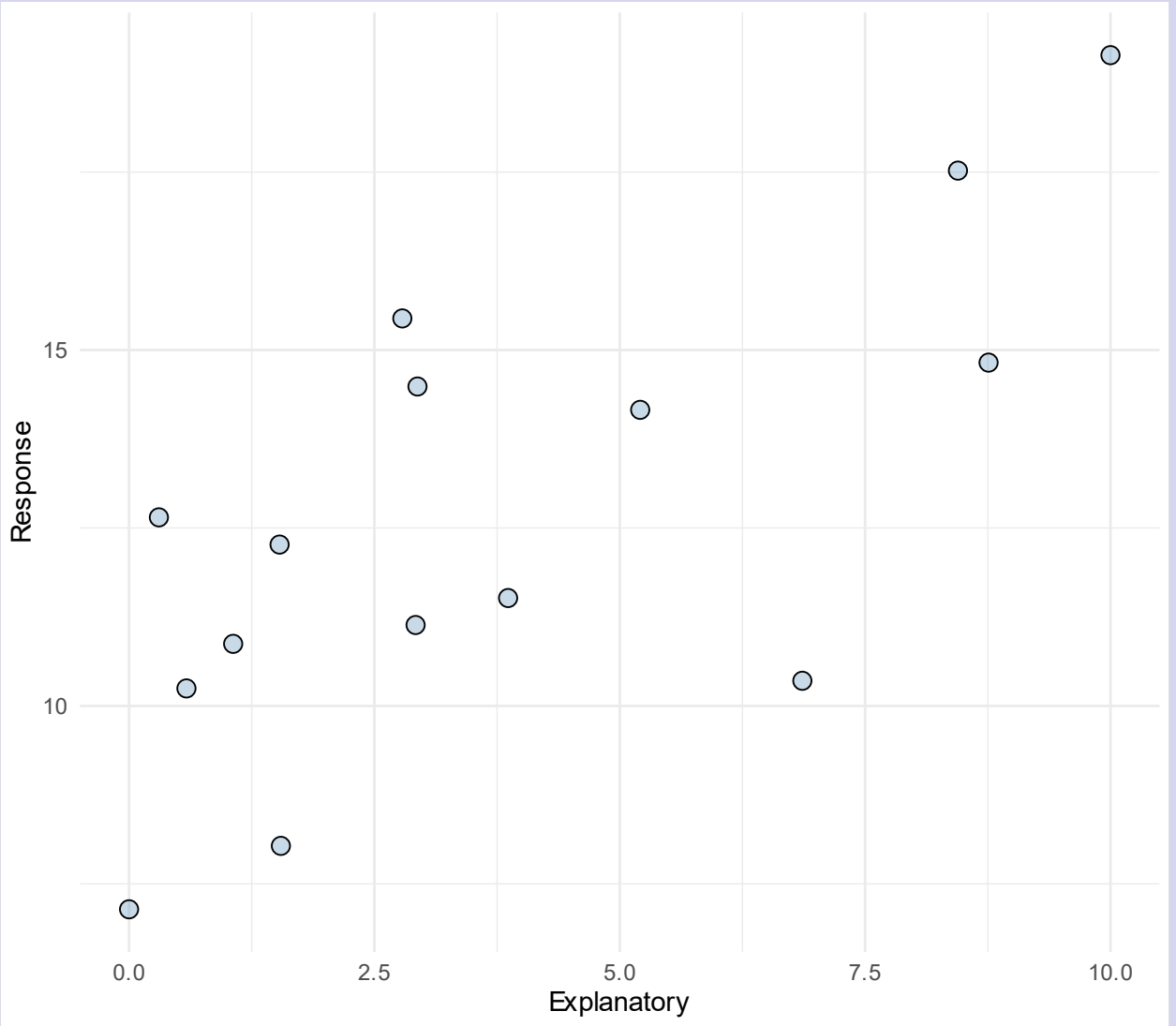
Correlation: Essential Concepts

What is correlation?

Correlation: Essential Concepts

Let's review covariance:

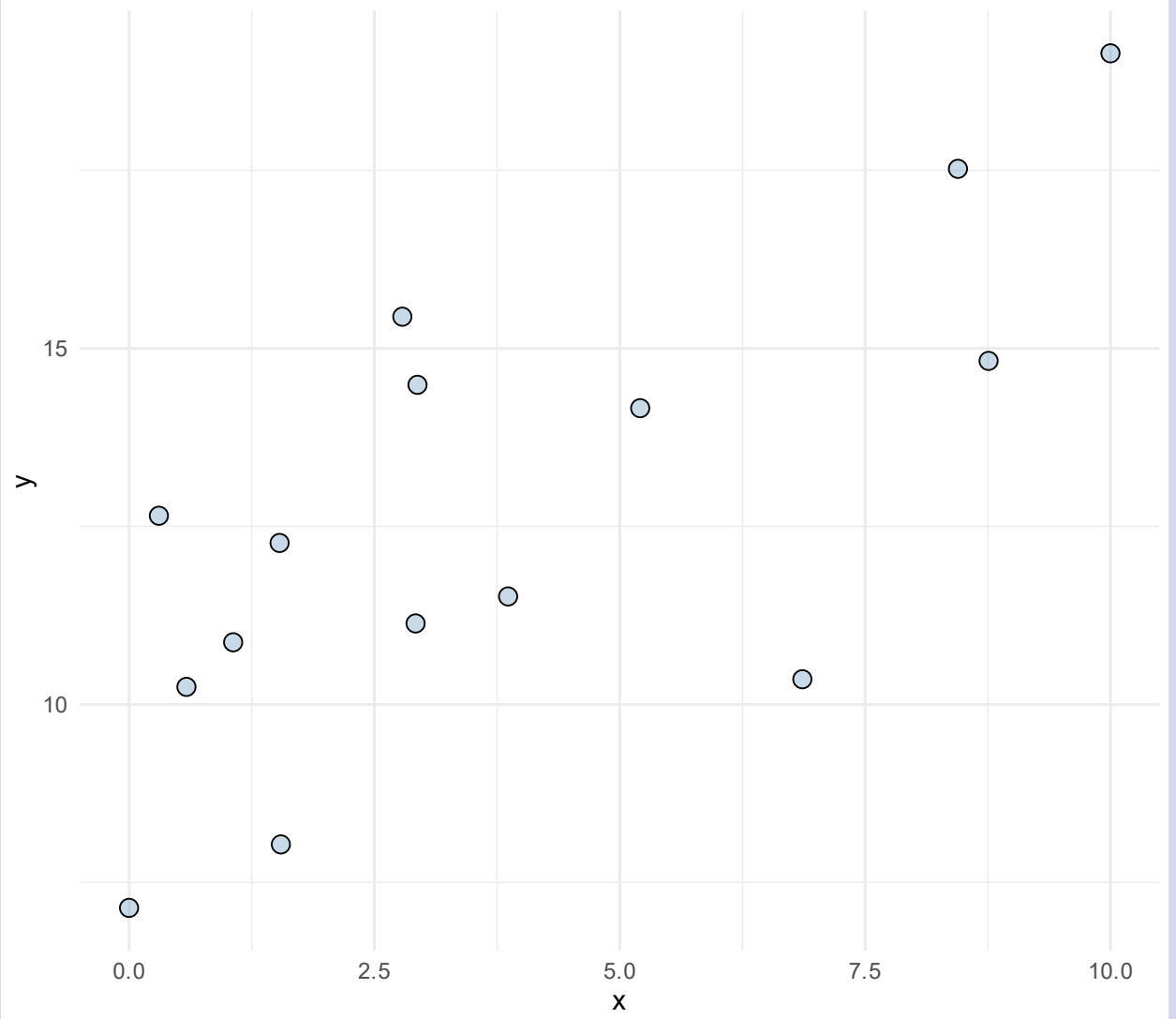
- Quantifies the strength of a **linear** or **monotonic** relationship between two continuous variables.
- Tells us the **strength** and **direction** of the relationship.
- Does **not** tell us about the **magnitude** of the relationship.



Correlation: Essential Concepts

Normalization: correlation vs. covariance

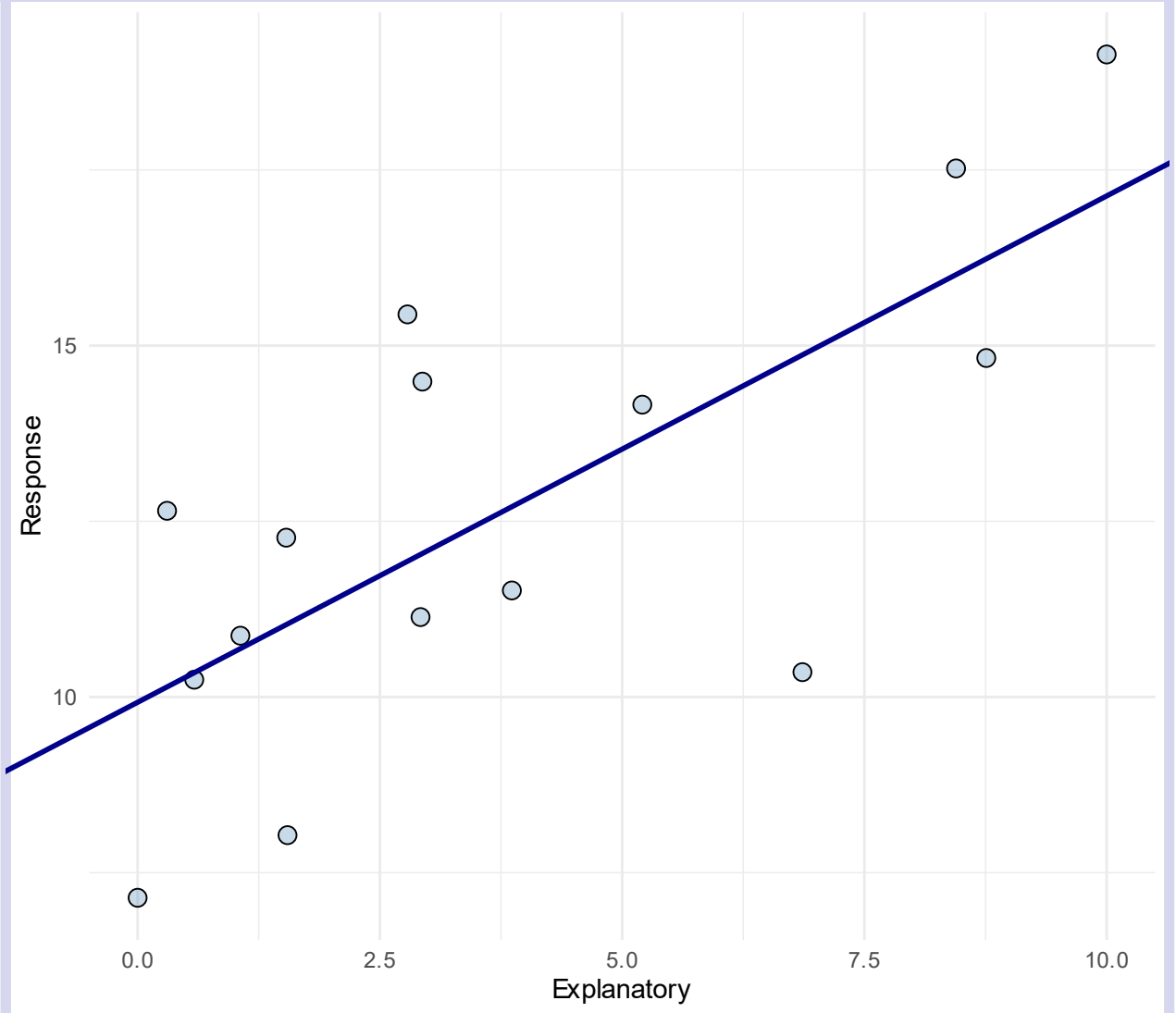
- Correlation is a normalized version of covariance.
- Covariance tells us how well coordinated two variables are
- The correlation coefficient, (r), **quantifies** how close are the values to the best fit line.
- Correlation analysis **does not** tell us anything about the line itself.



Correlation: Essential Concepts

Suppose the correlation coefficient is: 0.86, but how do we describe the magnitude of the association?

We want to get specific about the relationship between the two variables.

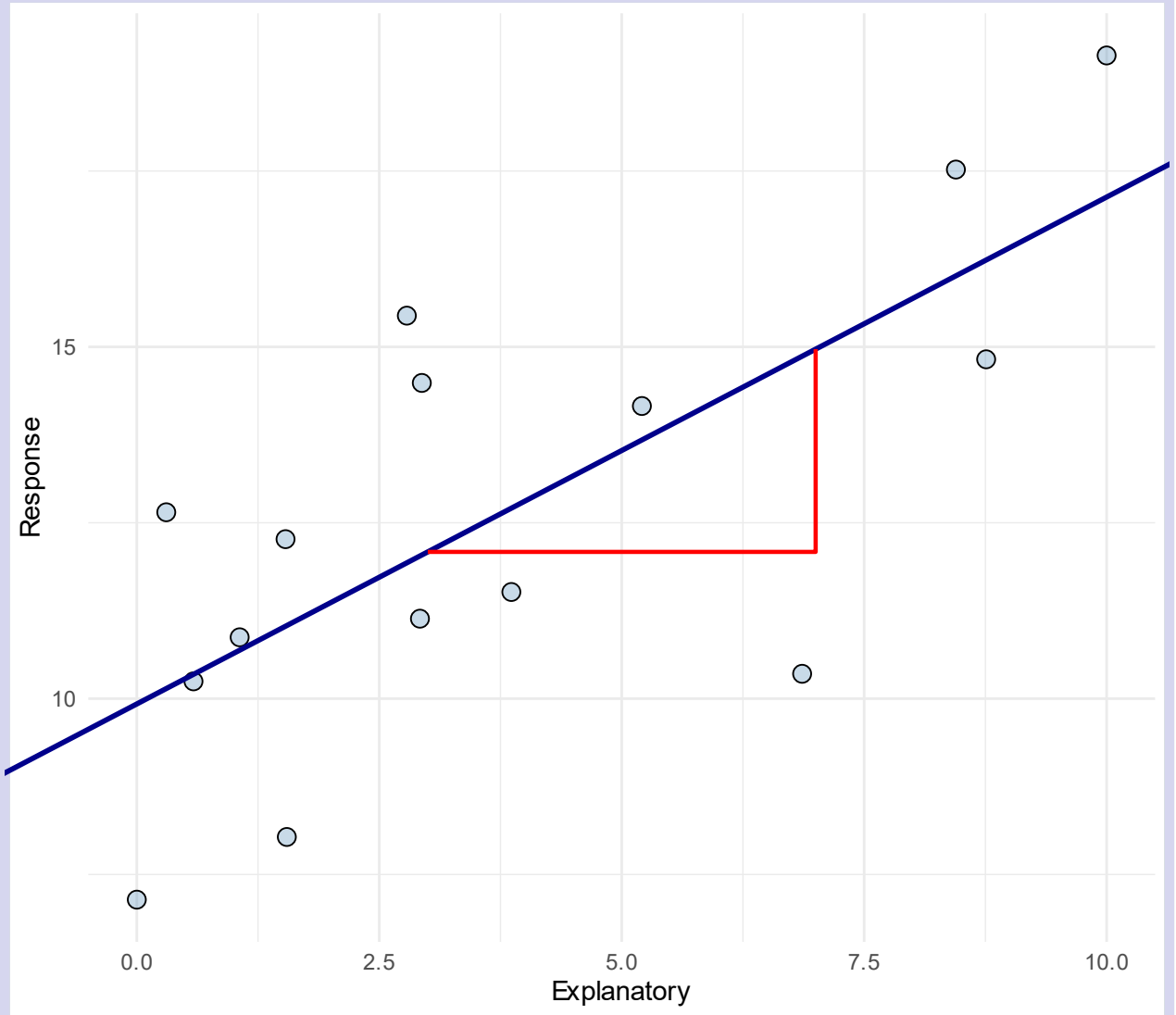


Correlation: Essential Concepts

Suppose the correlation coefficient is: 0.86, but how do we describe the magnitude of the association?

We want to get specific about the relationship between the two variables.

Linear Regression is a family of related techniques that we can use to learn about these relationships.



Green Spaces Survey

- Link in Moodle (or click on image)

Green Spaces

This survey/questionnaire attempts to get a better understanding of students perception of green spaces.
Responses to this survey are anonymous

Green Space: A green space is defined as publicly accessible areas with natural vegetation such as trees, plants, grass, etc. May include built environmental features (Structures, buildings, benches, etc).

Regression: Essential Concepts

What is regression?

What is linear regression?

Regression: Essential Concepts

What is regression?

- A modeling paradigm where we try to explain variation in one response variable using one or more predictor variables.

What is linear regression?

- A type of regression that makes specific assumptions about the form of the relationship between predictors and the response

Regression: Model Thinking

With **Regression**, our goal is to create a model to help us **understand** the relationships among variables in our data.

- Note that in real data, relationships don't have to be linear, or even monotonic.

We'll start with **linear regression** because it is the simplest type, and it often describes associations very well.

But remember: regression is just one type of modeling…

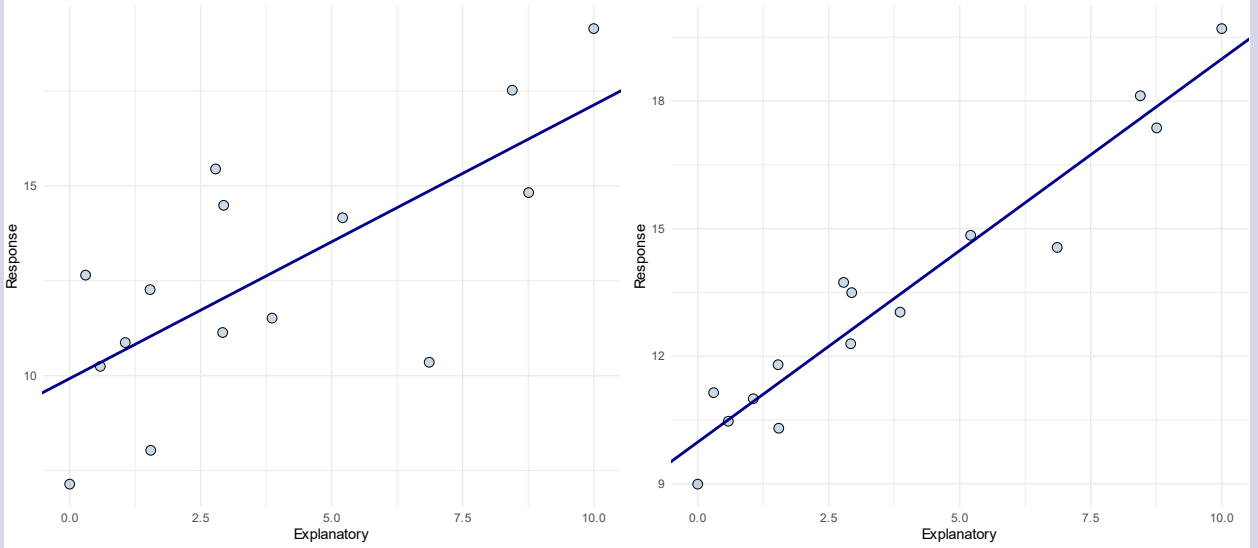
> Always remember that "All models are wrong, but some models are useful."

Our goal is to **fit a model** to our data, not the other way around!

Linear Regression: Filtting a Line

With a linear regression, we propose that the relationship between two variables can be [approximately] described by a line. This leads to a few questions:

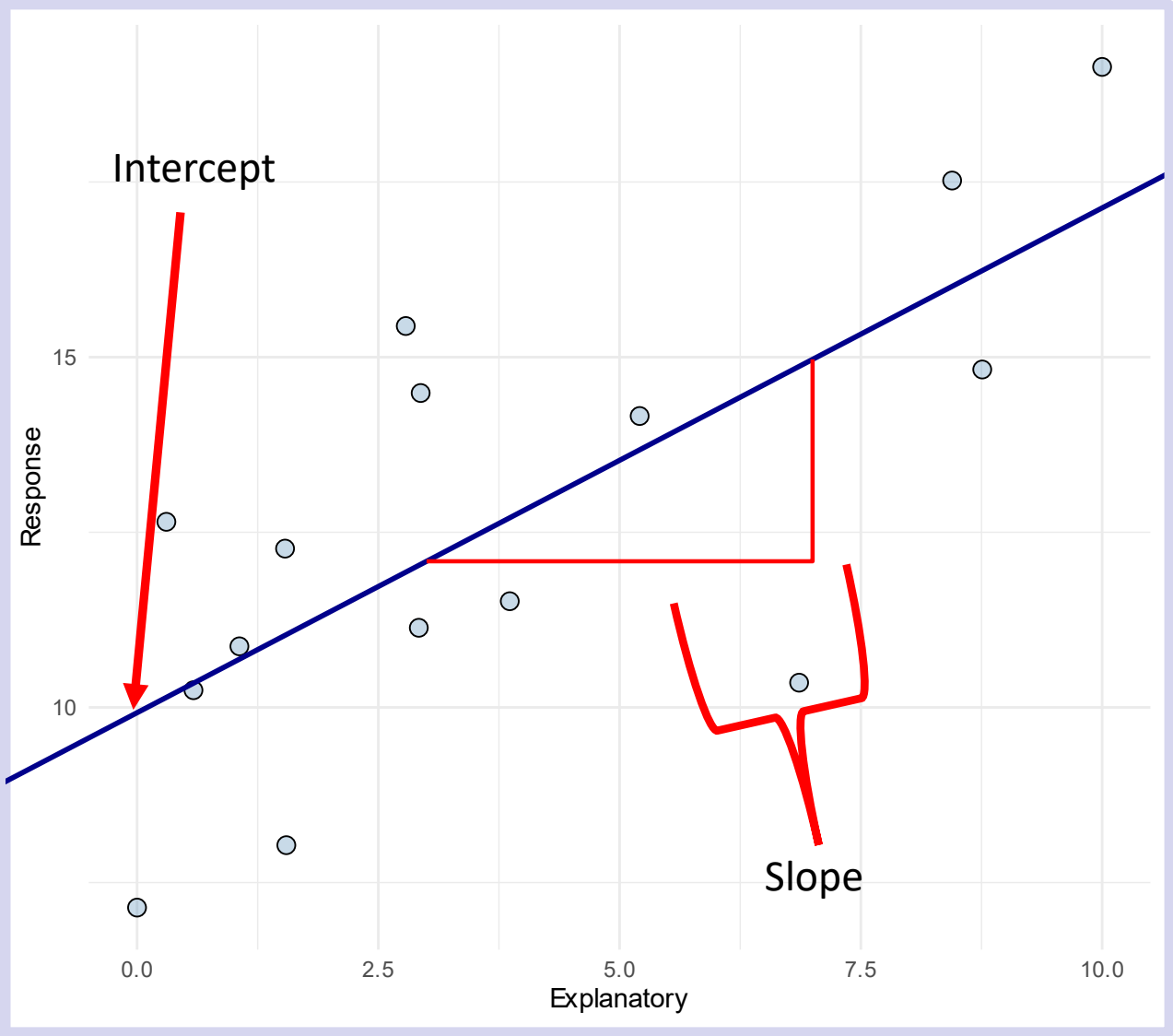
1. How do we know which line to choose?
2. What aspects of the line might be important?
3. How many **parameters** do we need to know?



Regression: Estimate Parameters

In a linear regression, our goal is to estimate the *parameters* of the best fit (straight) line:

- $y = a + bX + \varepsilon$
  - y: response variable
  - a: intercept
  - b: slope
  - X: explanatory variable
  - ε: error term



Regression: Estimating Parameters

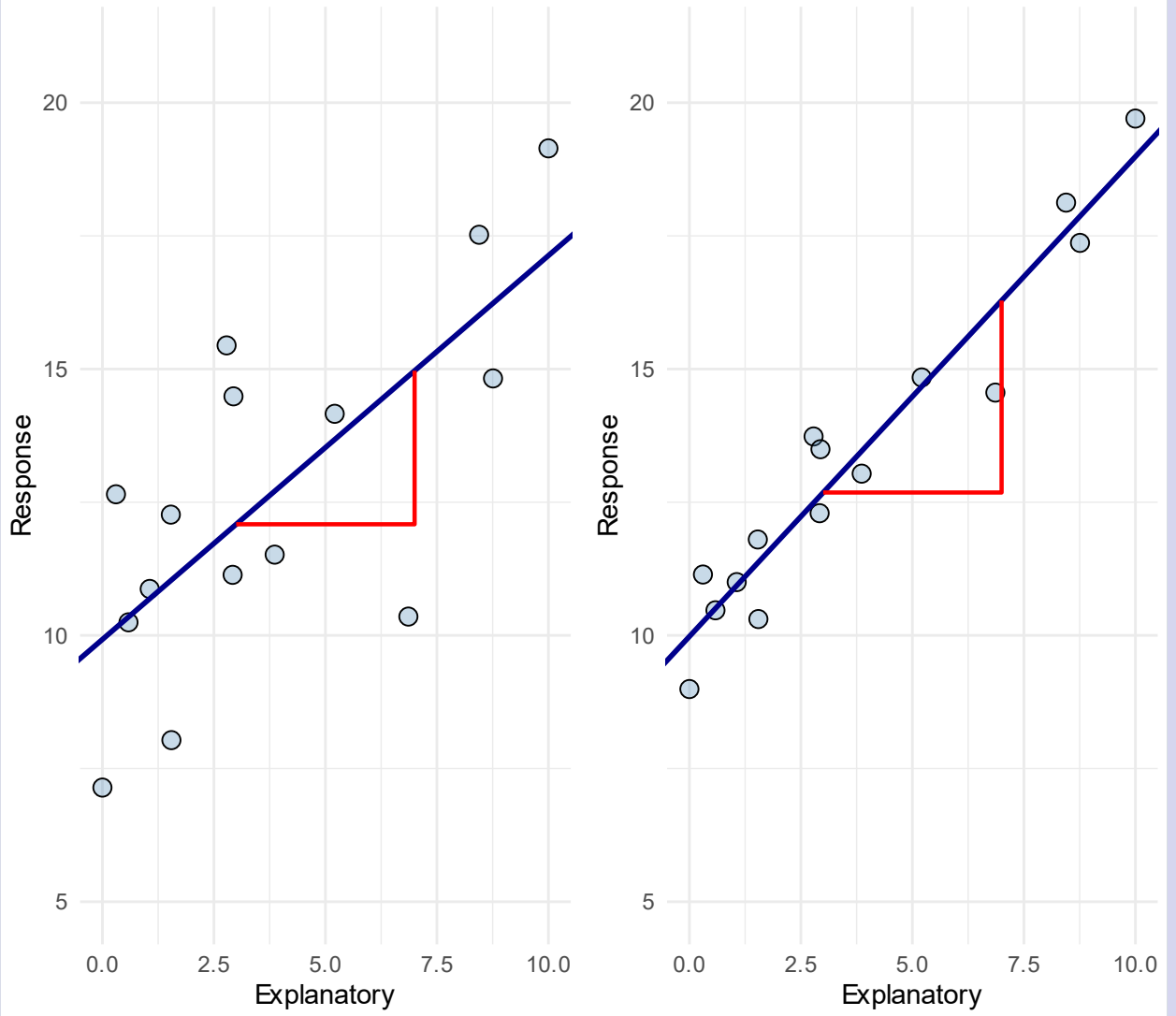
Estimated parameters of the best fit line $y = a + bX$ for these data:

Left plot:

- a: 9.91 (intercept)
- b: 0.72 (slope)

Right plot:

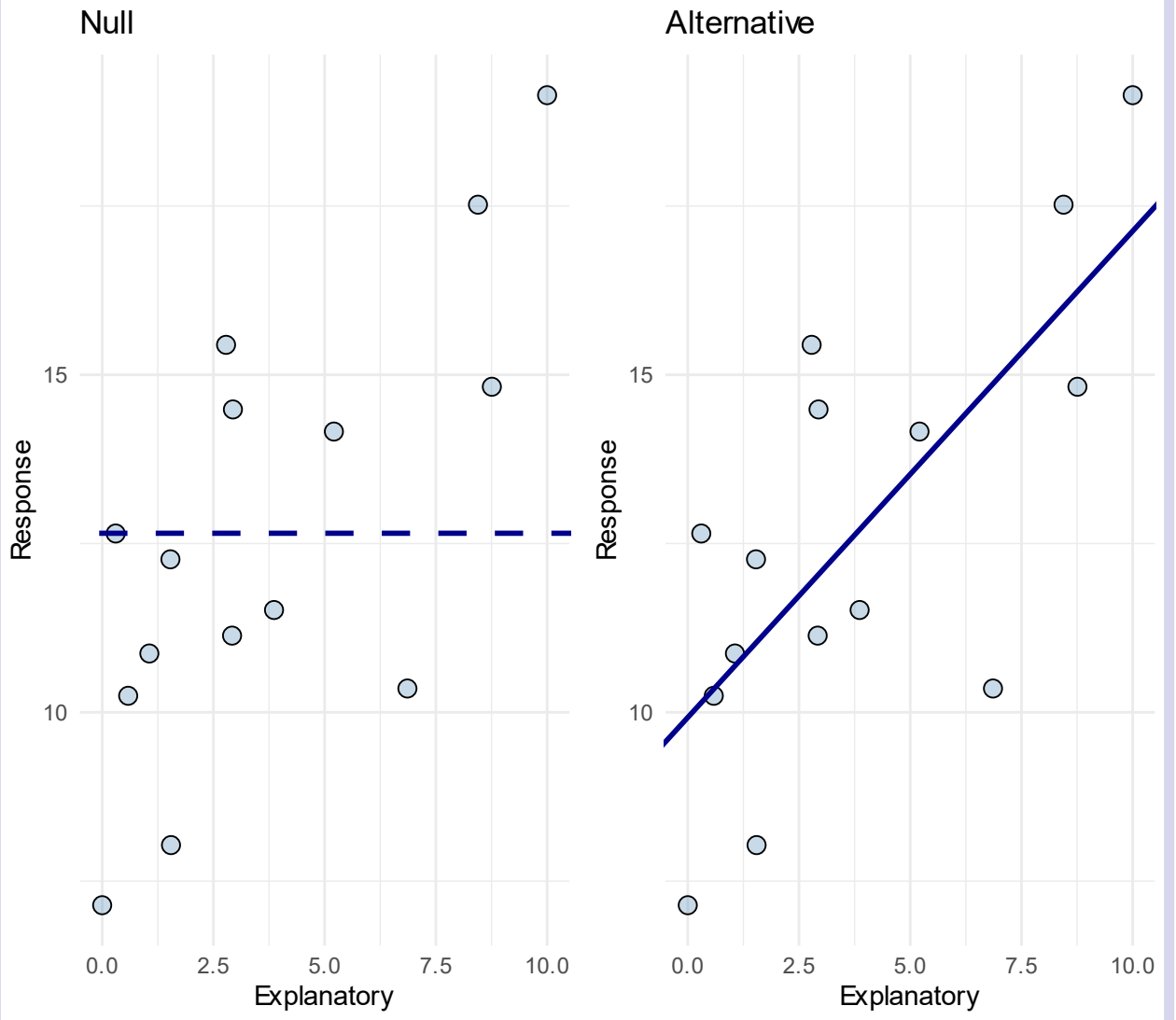
- a: 9.98 (intercept)
- b: 0.93 (slope)



Regression: Slope Inference

Regression analysis provides **inference** about the slope:

- *null* hypothesis: slope *is not* different from 0
- *alternative* hypothesis: slope *is* different from 0
- How might we **quantify** our evidence for a relationship?



Regression: Significance of Slopes

Regression analysis provides inference about the slope:

- *null* hypothesis: slope *is not* different from 0
- *alternative* hypothesis: slope *is* different from 0
- *p*-value of the slope

You can think of the p-value as:

1. What is the probability that I observe a slope *at least as large* as the observed value if the null hypothesis were true?
2. What is the probability that I observe a slope *at least as large* if the relationship between the predictor and response were completely random (no relationship)?

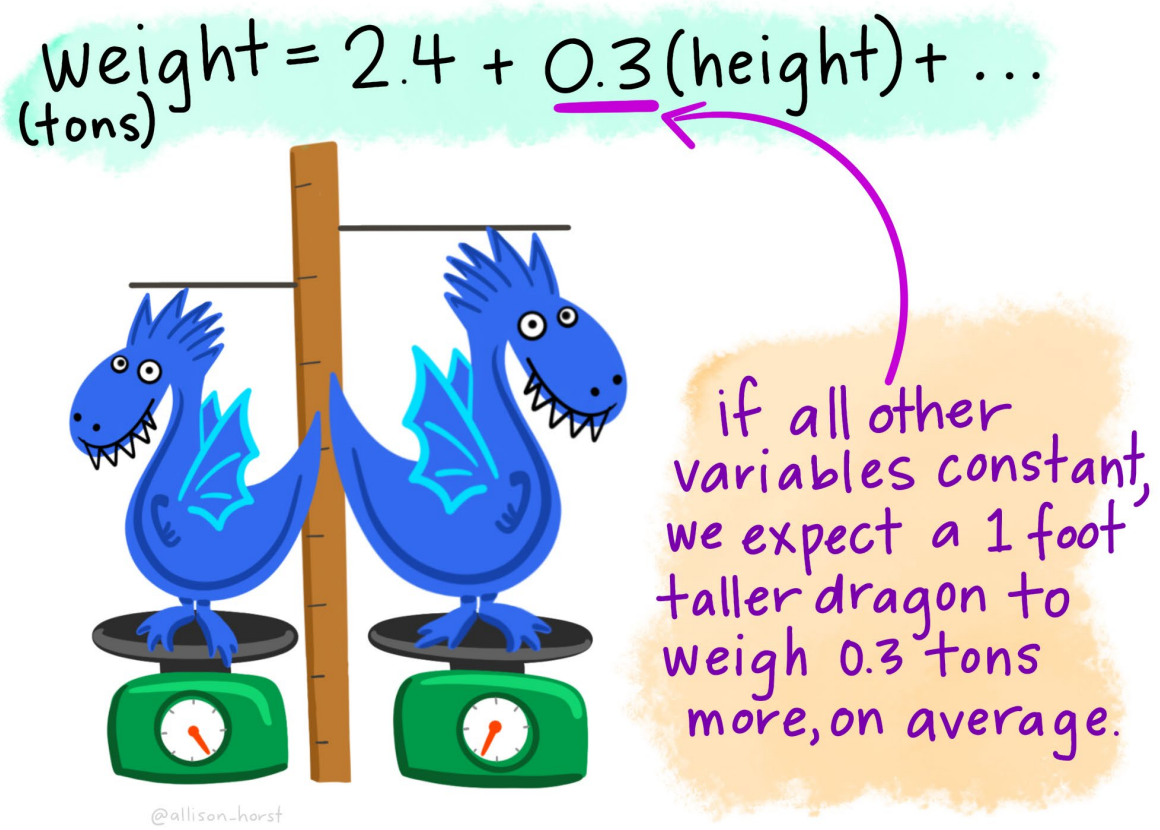
The Dual Model Paradigm

What is a Regression?

Regressions embody the dual-model concept

Regression is a modeling paradigm in which we specify a mathematical relationship between independent and dependent variables.

- A regression includes a *deterministic model* to specify the average behavior.
- It specifies a *stochastic model* to describe the variability around the average behavior.



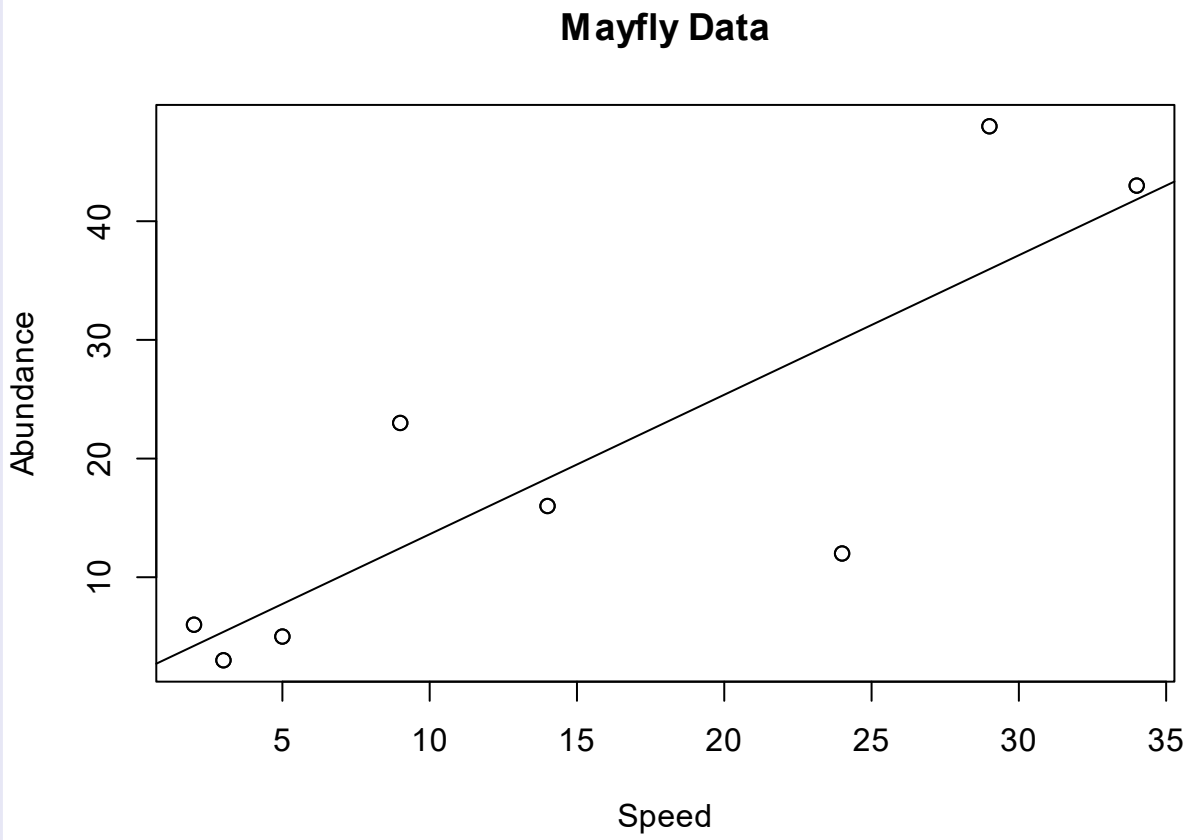
Artwork by @allison_horst

What is the Dual Model Paradigm?

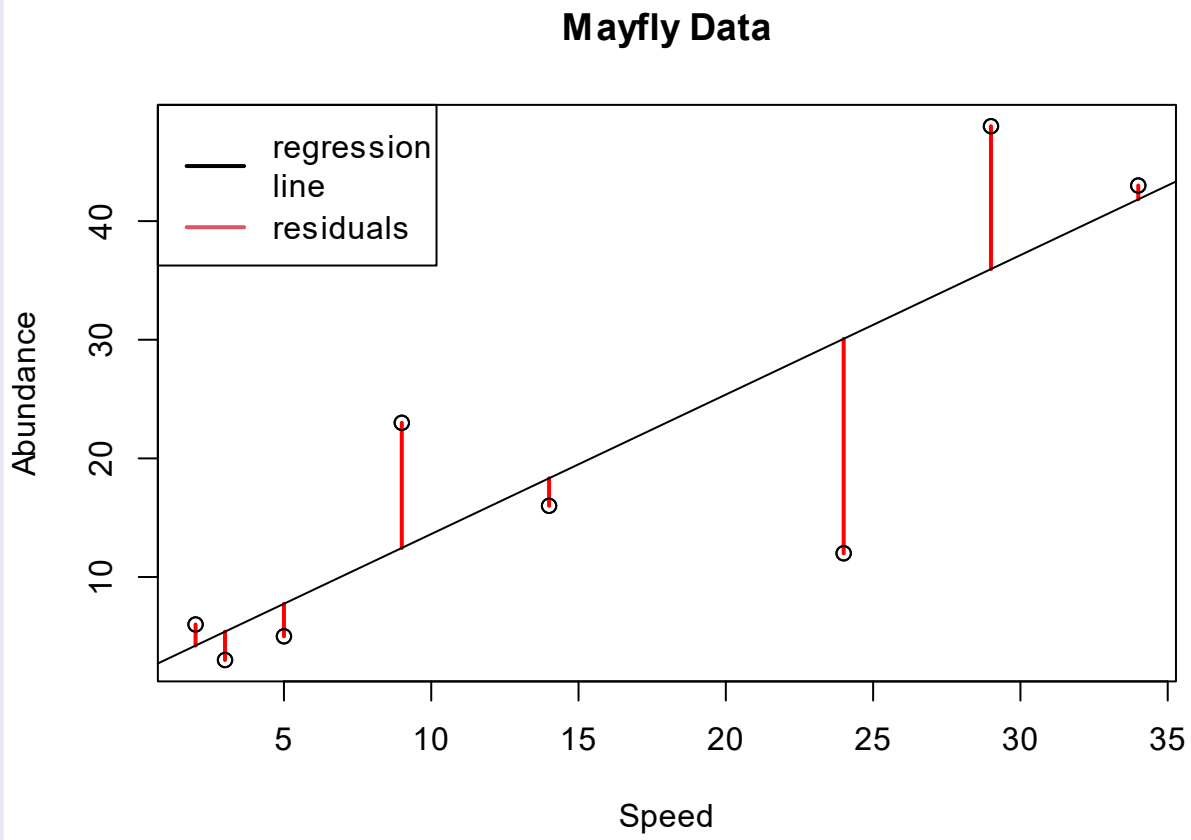
Deterministic Model	Stochastic Model
• The outcome of a deterministic process is always the same, there's no uncertainty. • We can use mathematical functions to model a deterministic process. • For example: a linear equation	• A stochastic process features uncertainty in its outcome. • Every *realization* of a stochastic process has a different outcome. • We can use a stochastic model to understand uncertainty. • Stochastic models are often described by probability distributions.

Dual Models

Deterministic Model: The Regression Line



Stochastic Model: The Residuals

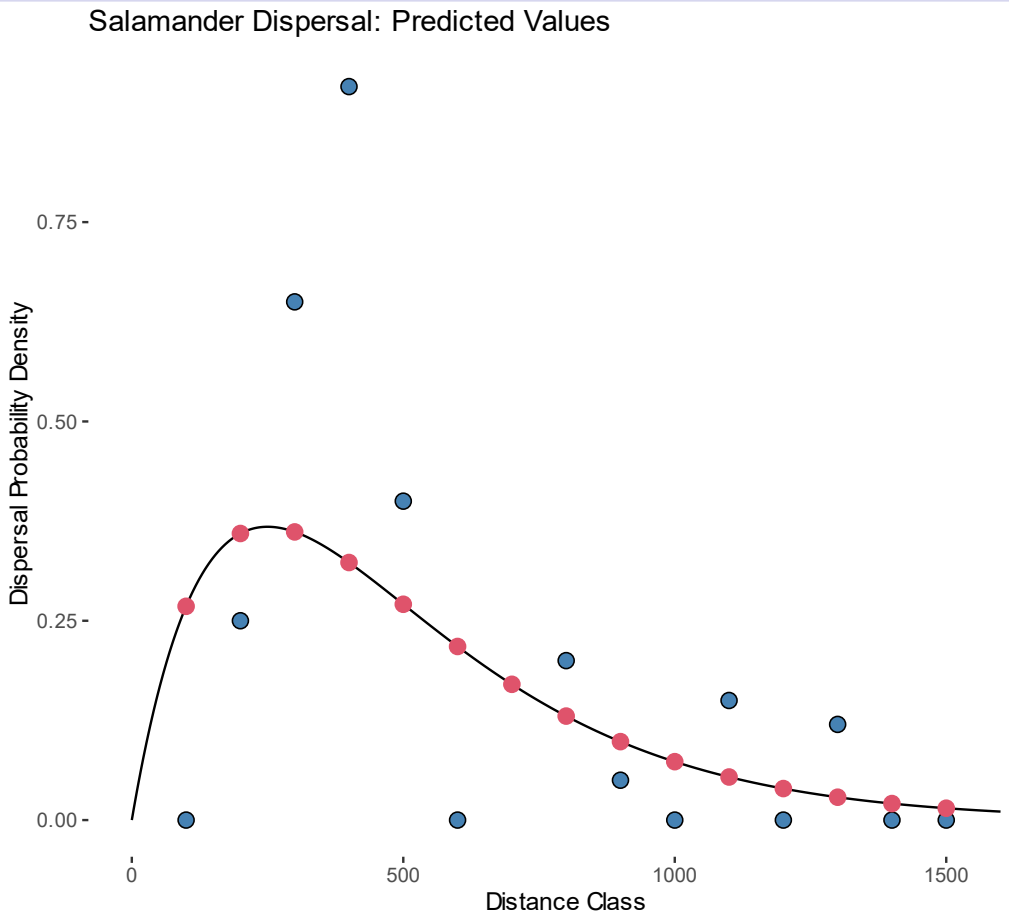


Dual Models

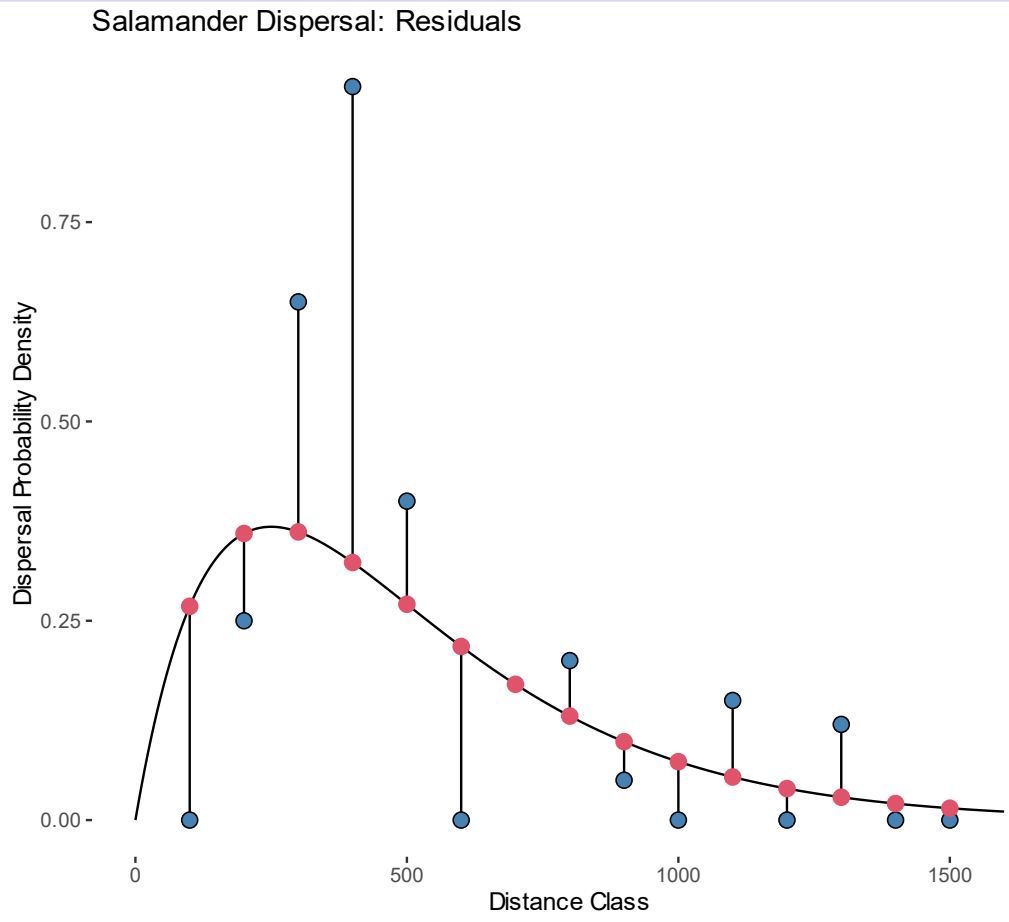
Deterministic Model: The Regression Line	**Stochastic Model: The Residuals**
The regression line describes the model's **predicted values**.	The residuals are the variation in the response that the model can't explain.
We don't expect that the observed values fall exactly on the regression line.	In a linear regression, we assume that the residuals are Normally distributed.

Model Residuals: Salamander Dispersal

Deterministic Model: The Predicted Values



Stochastic Model: The Residuals!



Linear Regression Assumptions

Those pesky assumptions!

Linear Regression Assumptions

Linear regression makes four key assumptions:

1. **Normality:** The model residuals are Normally distributed
2. **Constant variance:** The variance in the response variable is the same *for all values of the predictor(s)*. Also called homoscedasticity
3. **Independence:** Observations are independent
4. **Fixed x:** We have perfect precision in our measurement of predictor variables

A fifth requirement is **linearity:**

- The relationship between predictors and response is linear *in the parameters*

We also expect our response variable to be continuous.

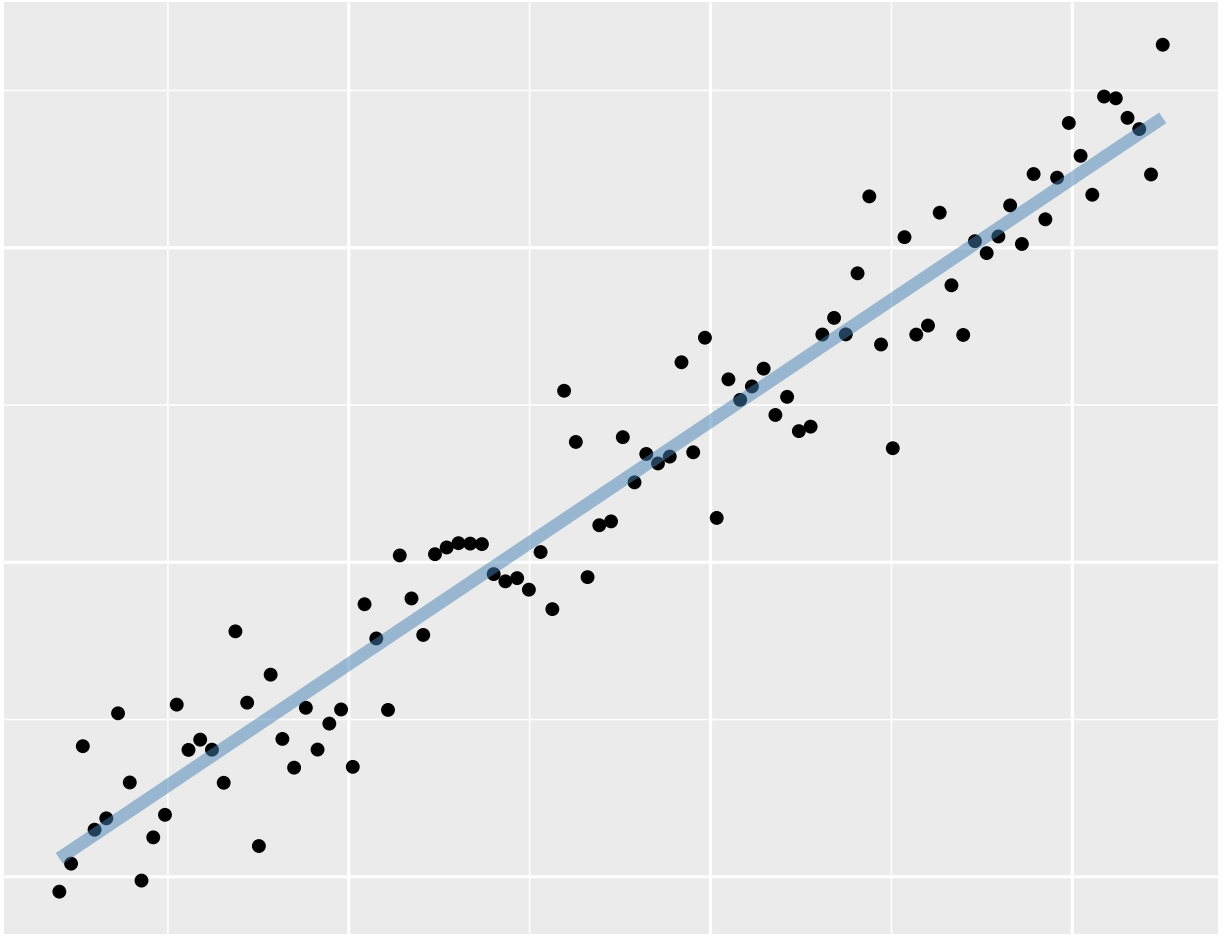
- Sometimes we can ignore this requirement

[Residual] Normality Assumption

- Under repeated sampling, data would be normally distributed *at each x*.
- Normally distributed around each *predicted value* in the *deterministic model*.
- This assumption is often misunderstood to mean that the values for each variable in a data set must be normally-distributed by themselves.
- But what is a residual?
  - The differenced between a predicted and observed value

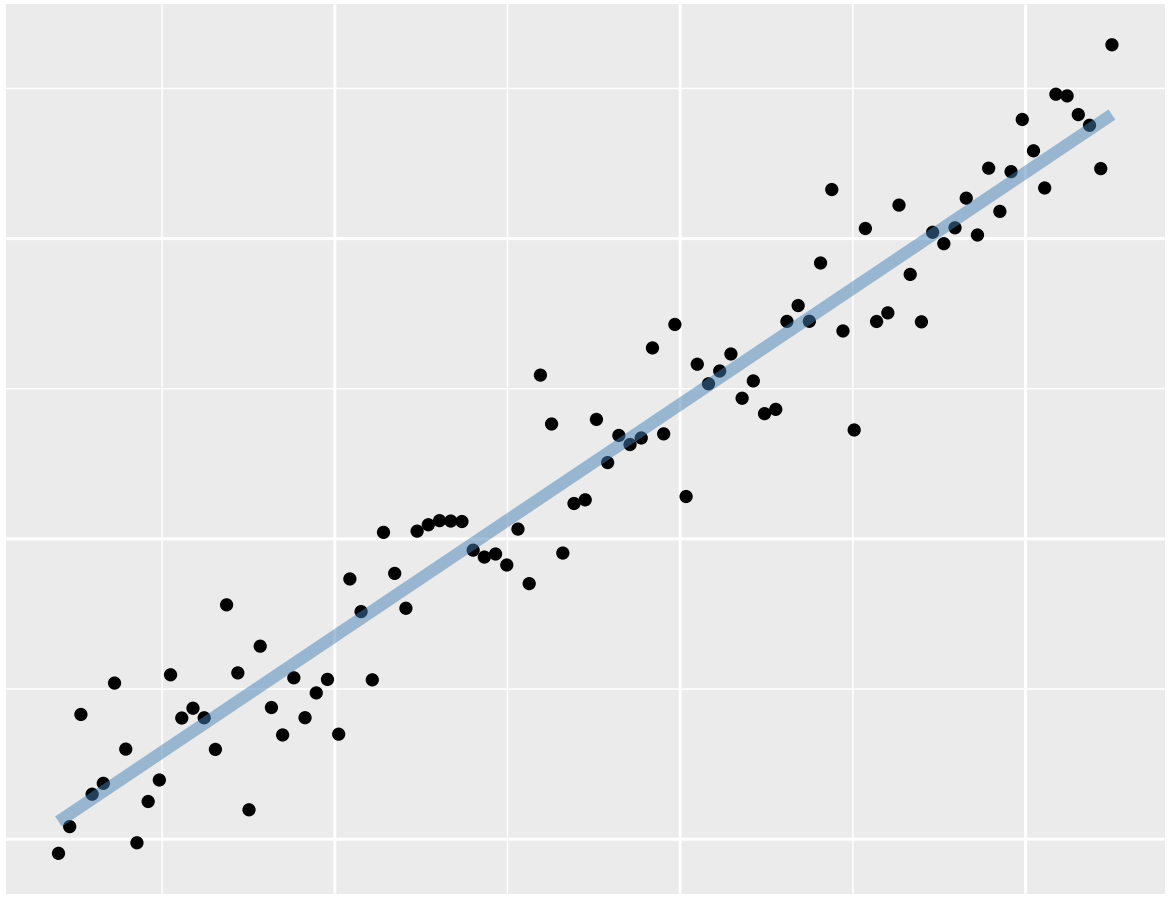
[Residual] Normality Assumption

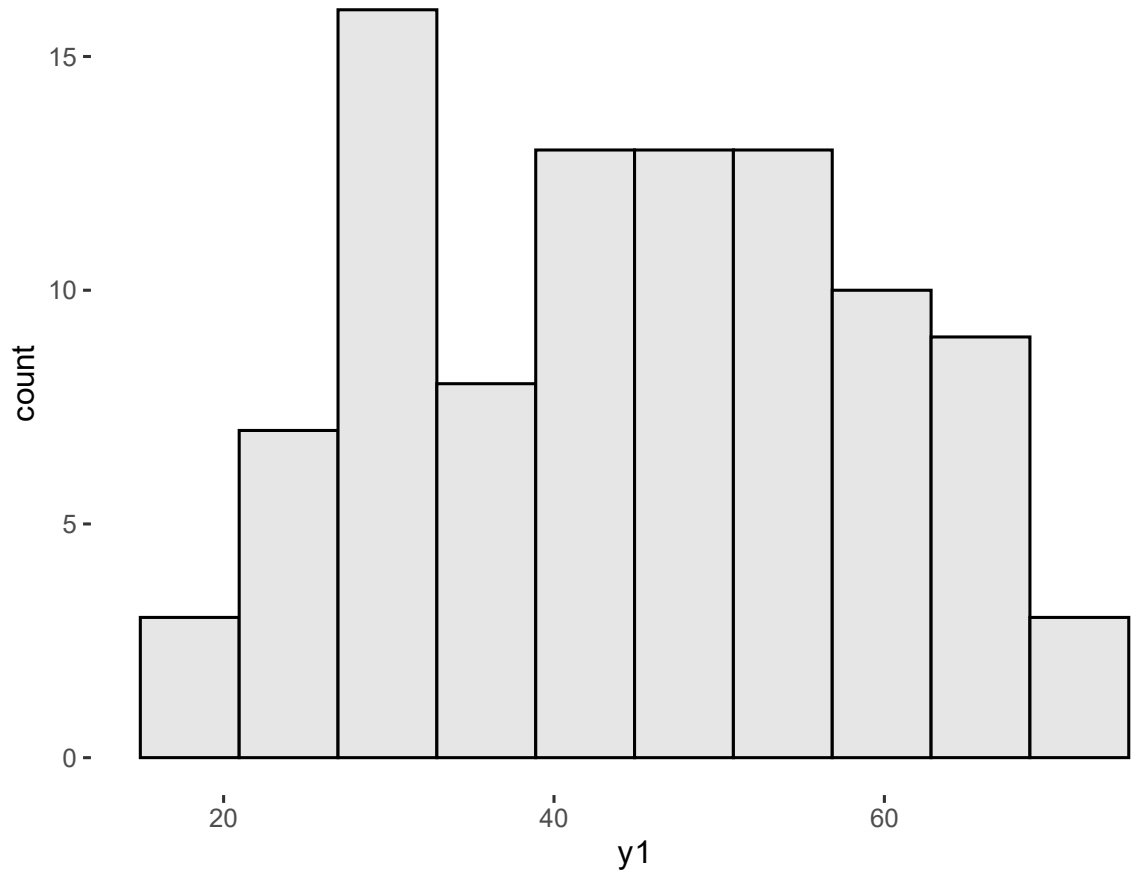
- Linear regression models assume that residuals are Normally distributed
- This does not mean that ‘the data are normally distributed’.
  - Usually, the data points themselves aren’t Normally distributed.
  - This is a frequent point of confusion.



Histogram of Raw Data

The following data look relatively well-behaved, however the histogram of the y-values suggests the distribution of values is pretty flat. A Shapiro test provides evidence of non-normality with p = 0.017.

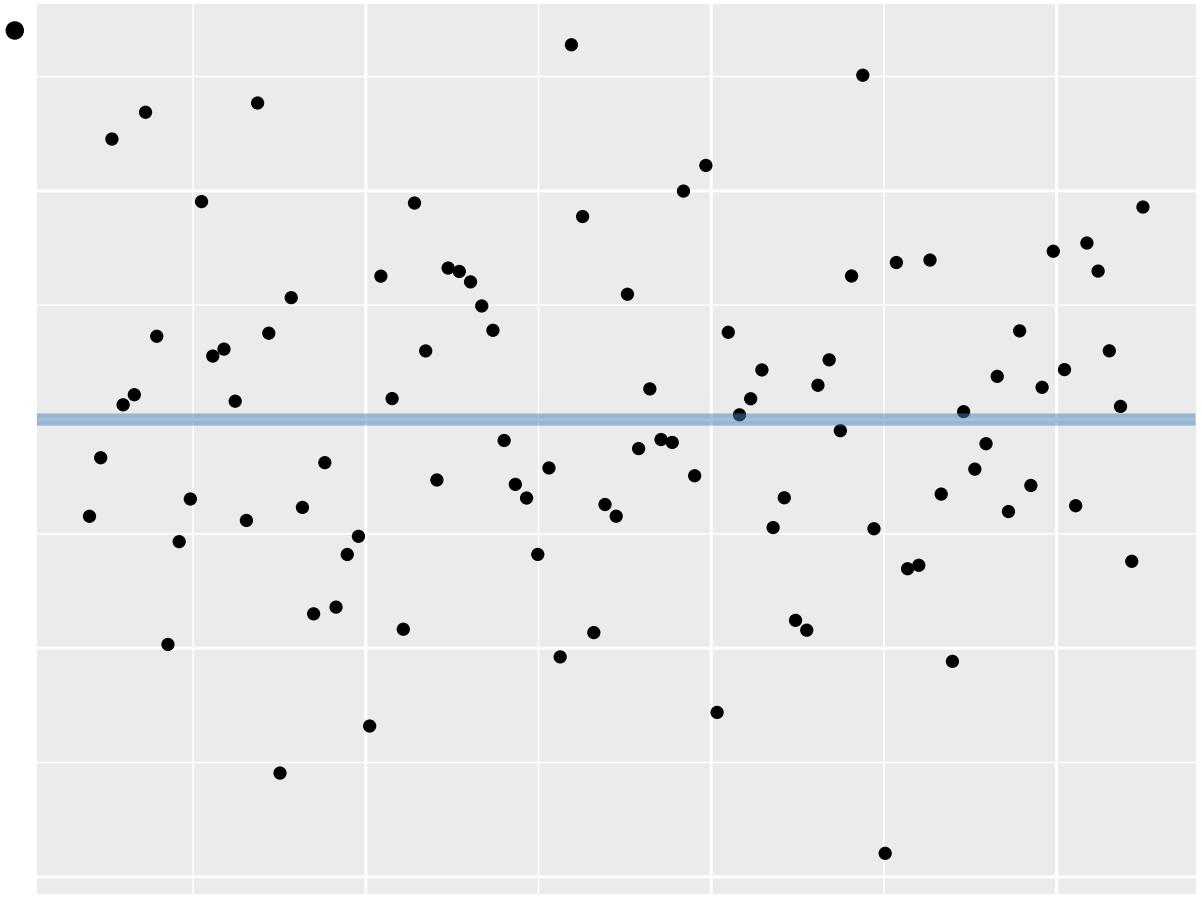


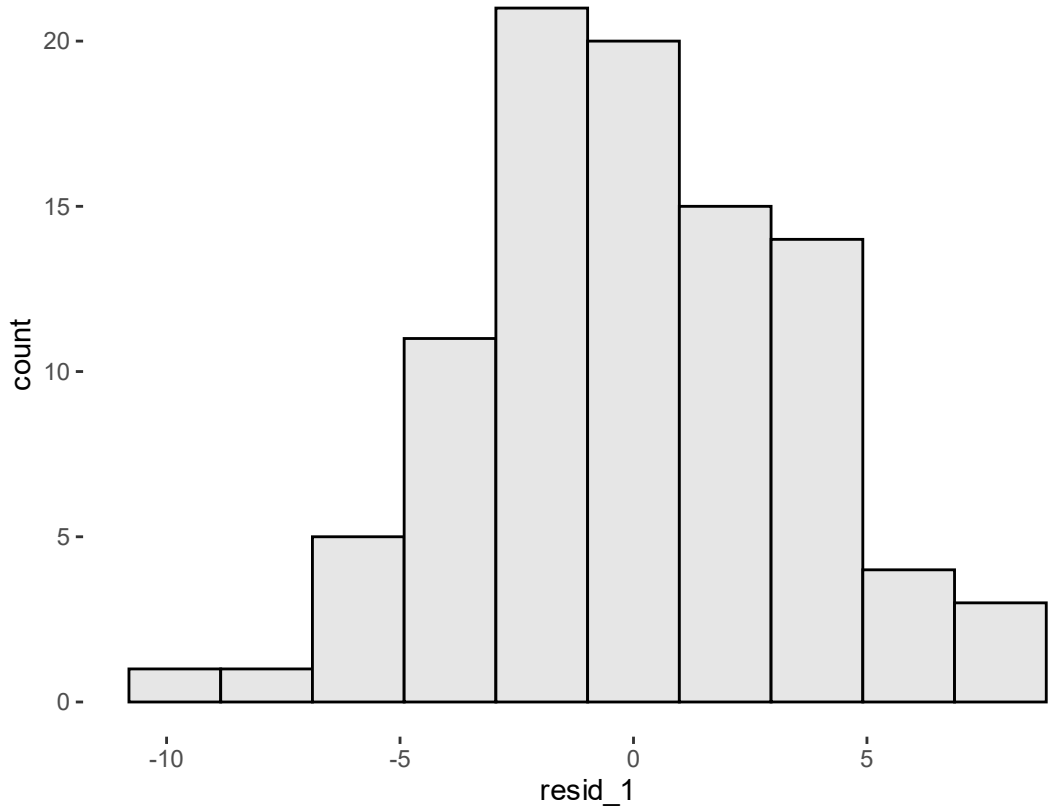


Histogram of Residuals

We really care about the normality of the *residuals* from a model.

A Shapiro test on the residuals suggests normality with p = 0.989.



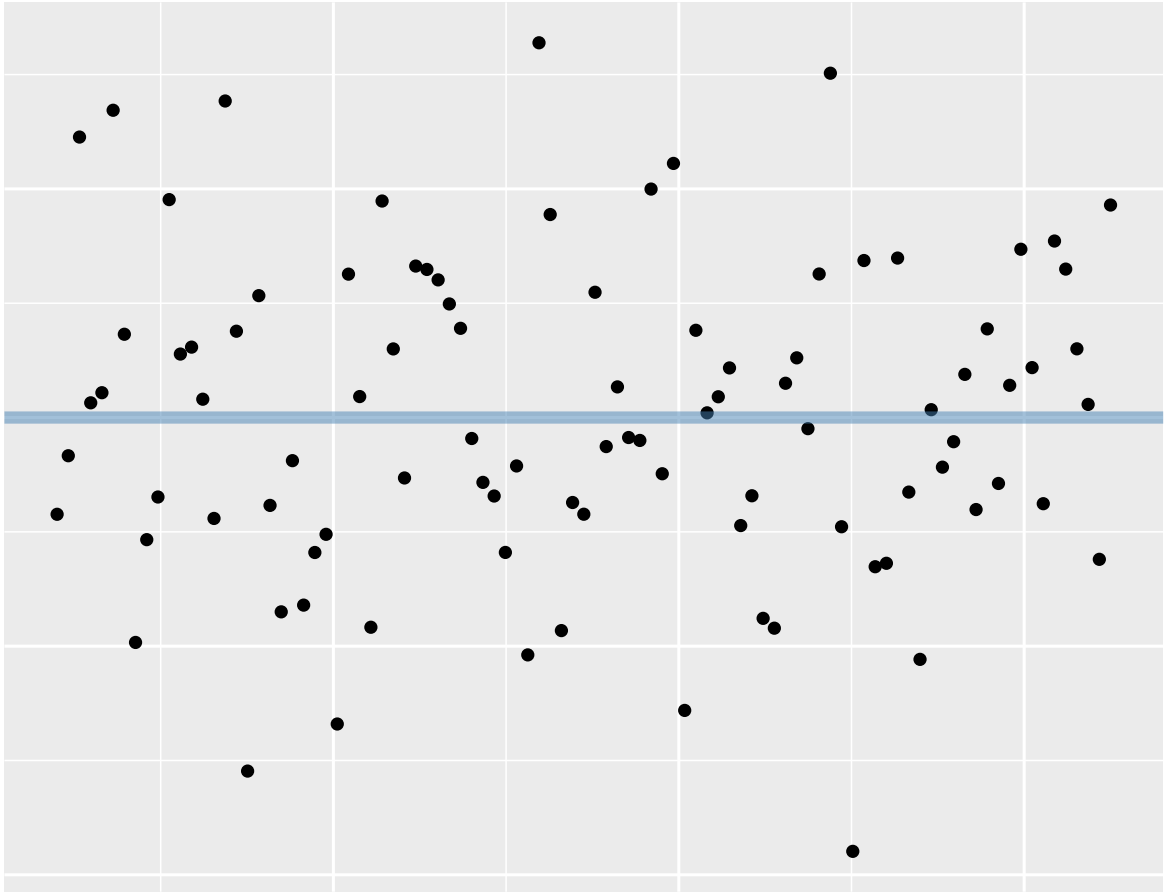


Homogeneity Assumption

The homogeneity assumption requires constant variance along the entire range of predictor values.

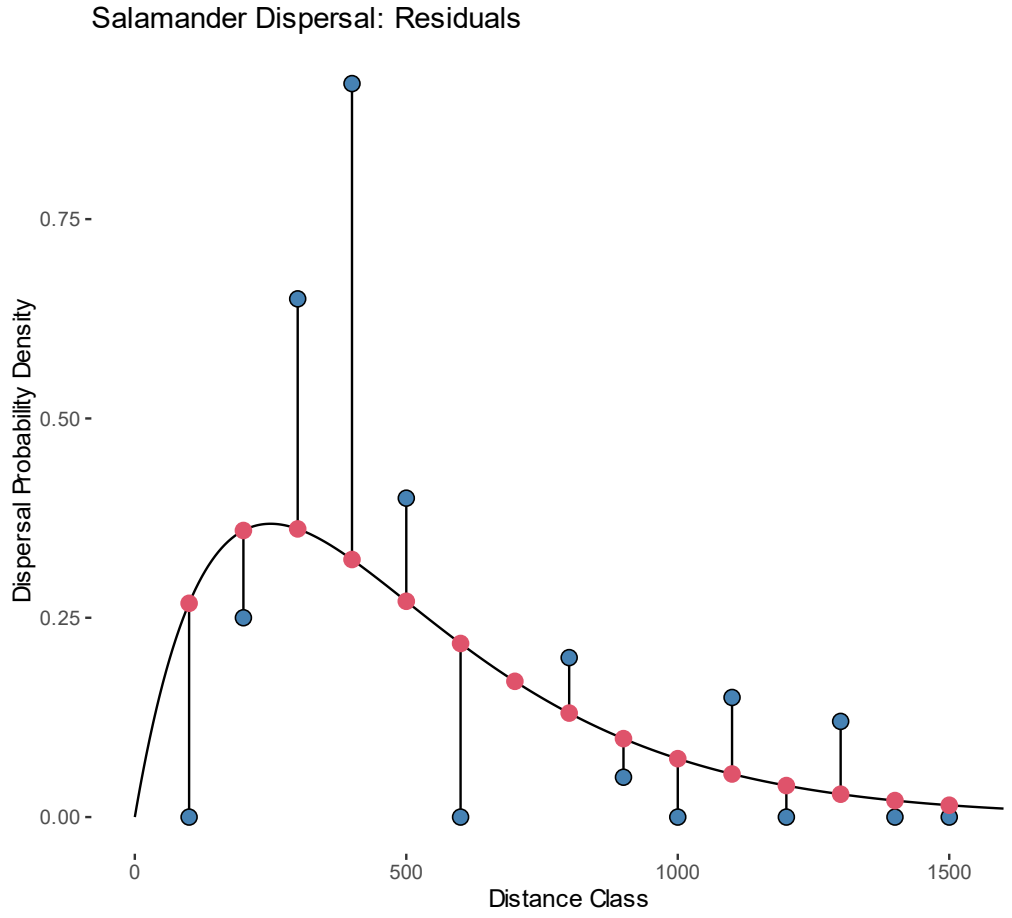
Key points of the assumption:

- The stochastic model is a Normal distribution.
- The spread parameter, σ is constant for all x.
- In other words, the variability does not depend on the value of x

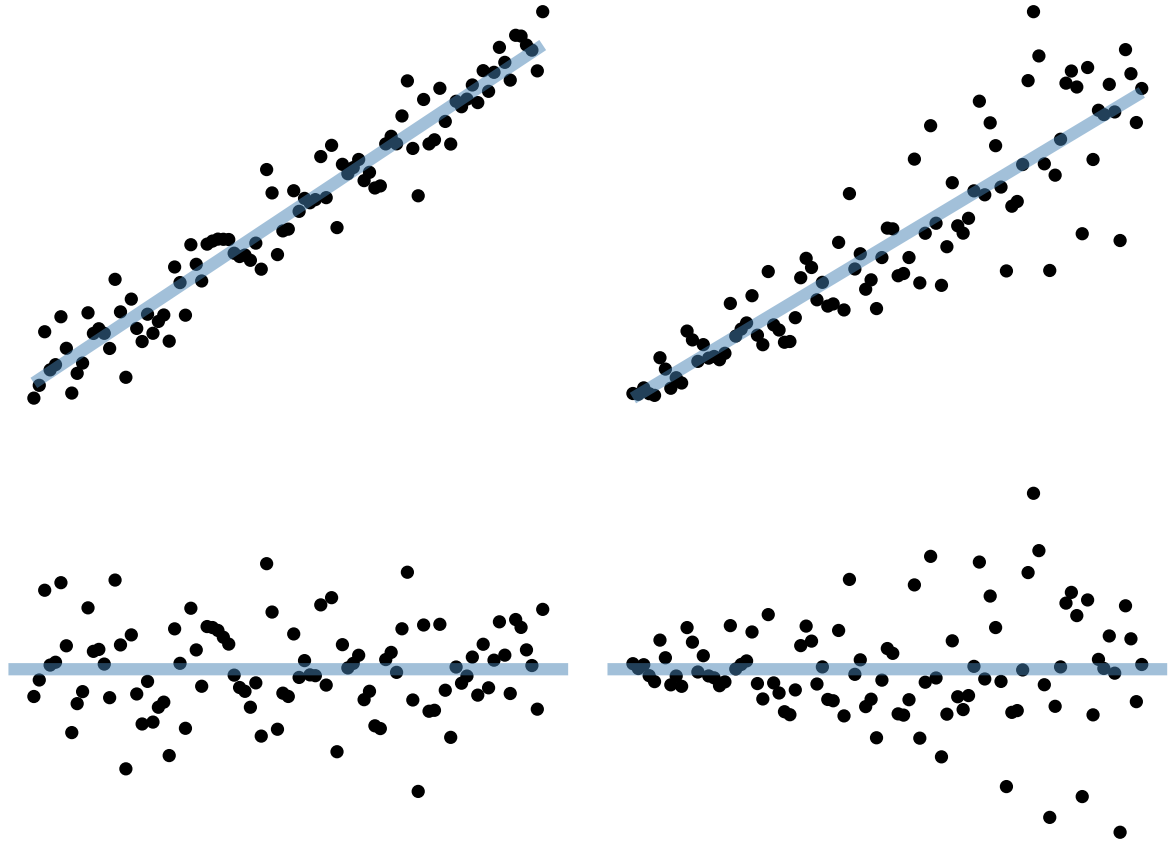


Homogeneity Assumption

Residuals of models of real data are often heterogeneous.



We don't like to see a megaphone shape



The Model Coefficient Table

How to Interpret Coefficients

Regression in pRactice: `lm()`

Remember our friend `lm()` from Analysis of Variance?

in algebra:

$$Response = a + b \times Explanatory$$

in `R` (a linear model):

```
Response ~ Explanatory
```

- Note that we don't specify the intercept

Suppose we have a `data.frame` called `df`:

```
head(df)
  Response Explanatory
1    27.22        2.88
2    35.94        7.88
3    15.94        4.09
4    33.06        8.83
5    35.97        9.40
6    17.50        0.46
```

Regression in pRactice: Building a Linear Model

We can use `lm()` to build a simple linear model using the formula notation.

The `summary()` function will print out a lot of helpful information about our model.

```
mod = lm(Response ~ Explanatory, data = df)
summary(mod)
```

```
summary(mod)


Call:
lm(formula = Response ~ Explanatory, data = df)


Residuals:
    Min      1Q  Median      3Q     Max
-8.1126 -1.6674  0.2598  2.7585  5.9932


Coefficients:
            Estimate Std. Error t value Pr(>|t|)
(Intercept)  14.5011     3.1439   4.612  0.00173 **
Explanatory   2.3353     0.4896   4.770  0.00141 **
```

Model Coefficient Interpretation

What do the numbers in the coefficient table mean?

Suppose we had a model of Total Length ~ SVL with coefficients:

Slope = 1.1, intercept = -2 with units of millimeters

Intercept: this is the value of the response when all of the predictors equal zero. In English:

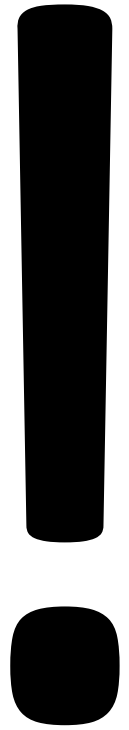
"A salamander with a SVL of zero would have total length of -2 millimeters"

Slope: For every 1-unit increase in x, the slope tells us how large the increase in the response should be, on average. In English:

"For every 1-millimeter increase in SVL, we expect a 1.1-millimeter increase in total salamander body length."

Intercept Caveat

- The intercept is the value of the response when all predictor variables are set to zero.
- The intercept is often considered a “tuning parameter”
- We usually don’t observe predictor values in the range of zero.
  - The intercept is then an **extrapolated** value.

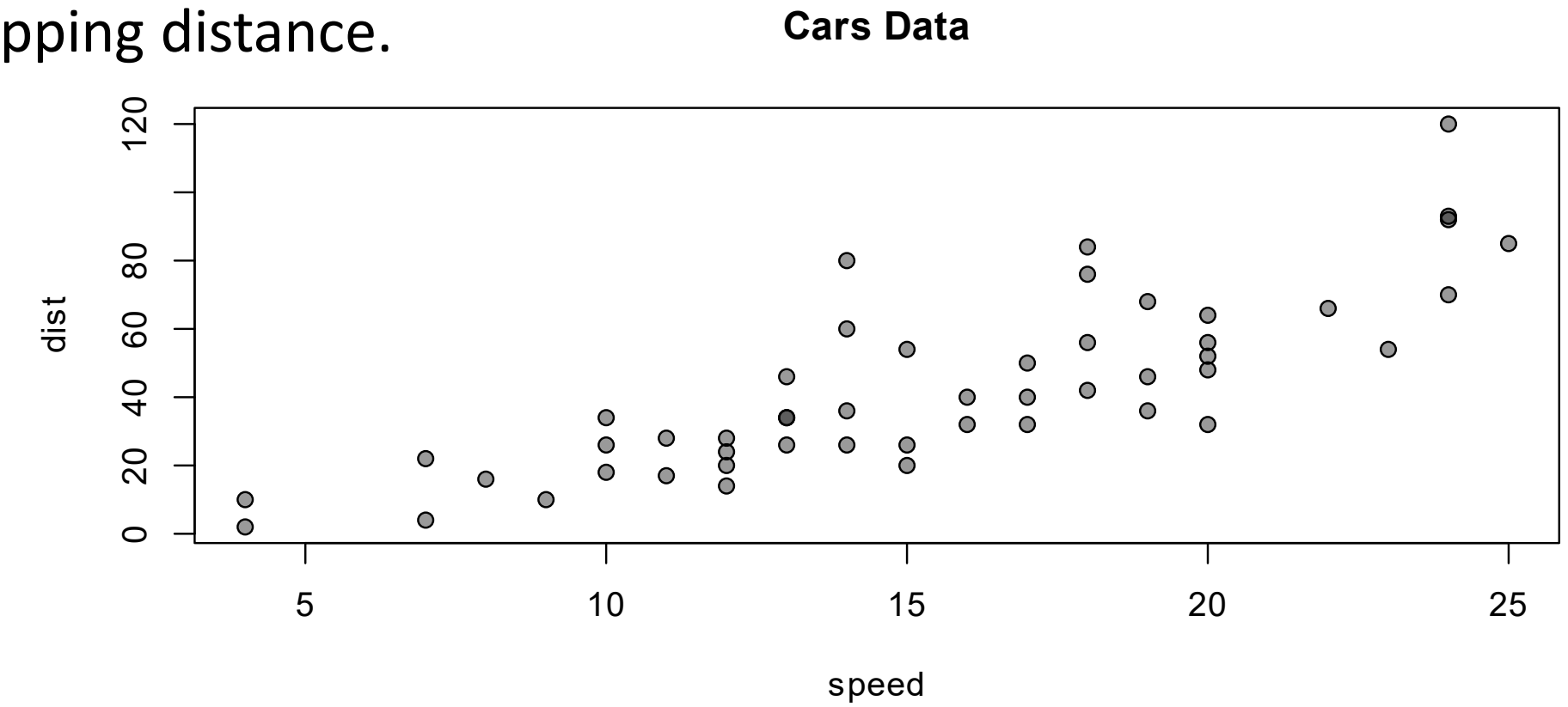


Simple Linear Regression: Example

The Cars Dataset

Example simple regression: Cars data

- The `cars` data set contains speed and stopping distance observations from 50 trials.
- The goal was to quantify the relationship between driving speed and stopping distance.



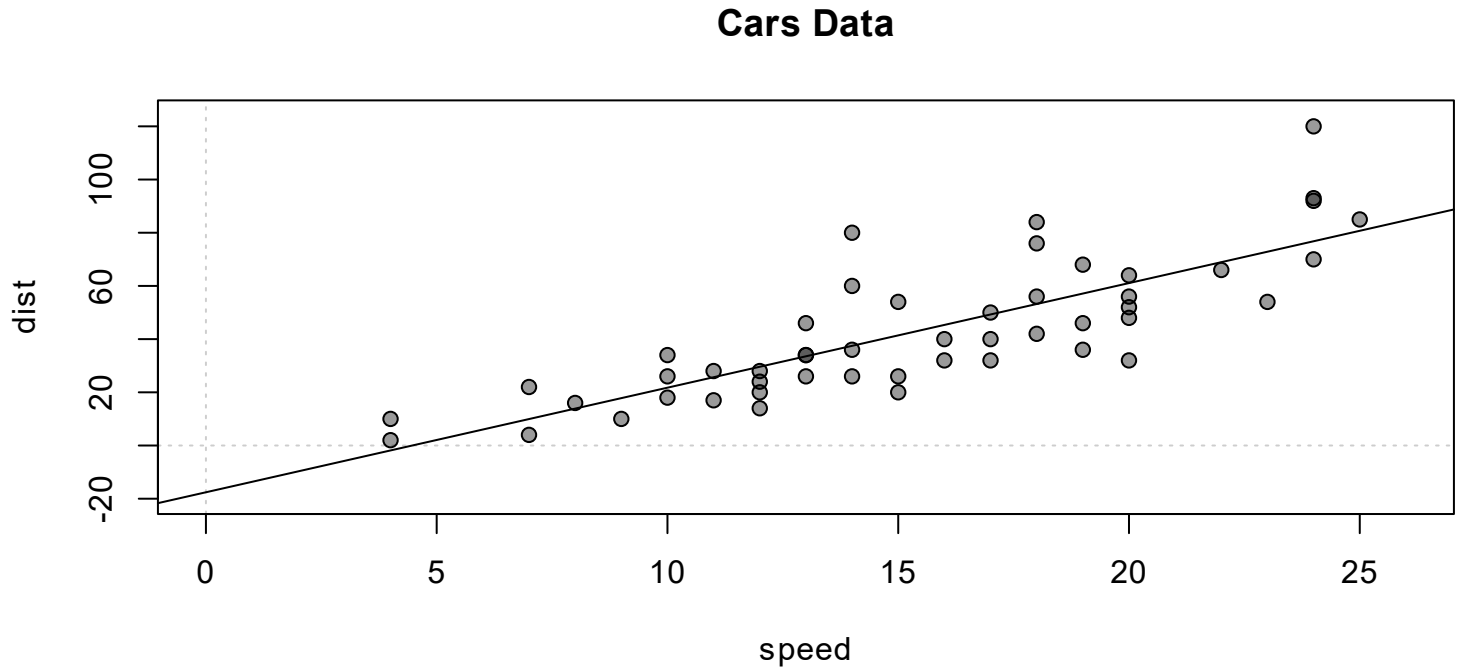
Cars: Fitting a Model

- Which variable is the predictor?
- Which is the response?
- We can fit a simple linear model:

```
fit_cars = lm(dist ~ speed, data = cars)
```

```
plot(dist ~ speed, data = cars)
abline(fit_cars)
```

We can use abline() to plot the regression line.



Cars Model: What can we learn from the coefficient table?

```
Call:
lm(formula = dist ~ speed, data = cars)

Residuals:
    Min      1Q  Median      3Q     Max
-29.069  -9.525  -2.272   9.215  43.201

Coefficients:
            Estimate Std. Error t value Pr(>|t|)
(Intercept) -17.5791     6.7584  -2.601   0.0123 *
speed         3.9324     0.4155   9.464 1.49e-12 ***
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 15.38 on 48 degrees of freedom
Multiple R-squared:  0.6511,    Adjusted R-squared:  0.6438
F-statistic: 89.57 on 1 and 48 DF,  p-value: 1.49e-12
```

Model Formula

Residuals summary – we'll learn more about this later

Model Coefficients

Amount of variation explained by the model

Example simple regression: Cars data

Let's translate the model coefficients to English. You can use the `coefficients()` function in R to retrieve just the regression coefficients, without all of the information from the full summary:

```
coefficients(fit_cars)
(Intercept)       speed
  -17.579095    3.932409
```

- The intercept is about -17.6.
  - What does that mean, is it sensible?
  - How could you translate the speed coefficient (3.9) into an English sentence?
- The R2 was 64%: That means the model explains about 64% of the variation in the response.
  - Is that a lot? Is it a good model? What is the other 36%?

Example simple regression: Cars data

Let's translate the model coefficients to English. You can use the `coefficients()` function in R to retrieve just the regression coefficients, without all of the information from the full summary:

```
coefficients(fit_cars)
```

```
(Intercept)       speed
 -17.579095    3.932409
```

- The intercept is about -17.6.
  - What does that mean, is it sensible?
- How could you translate the speed coefficient (3.9) into an English sentence?
  - "For each 1-mph increase in speed, it takes about 4 additional feet to stop".

Cars Model: Regression Equation

- We can build the regression equation from the model coefficient output from R:

$$distance = -17.6 + 3.9 \times speed$$

- How does this help us?
- One way we can use regression equations is for **prediction**.

- Given our regression equation
- $distance = -17.6 + 3.9 \times speed$
- we could calculate an **expected** stopping distance for any possible speed.

Cars Model: Regression Equation

We could do the calculation by hand, but fortunately R has a built-in function to use a model fit object to obtain predicted values.

- Perhaps counter intuitively, this function is called `predict()`.
- `predict` expects a data frame with a columns for each of the model predictors. In the cars model we had only one predictor: `speed`.
- We could calculate the expected stopping distance for a car travelling at 34 mph:

```
predict(fit_cars, newdata = data.frame(speed = 34))
       1
116.1228
```

Regression Equation: Driving at 450 mph

What about a car traveling at 450 mph?

```
predict(fit_cars, newdata = data.frame(speed = 450))
       1
1752.005
```

The equation predicts that you would need about 1/3 mile to stop.

Regression Equation: Stopped Car

A stopped car (0 mph) would need:

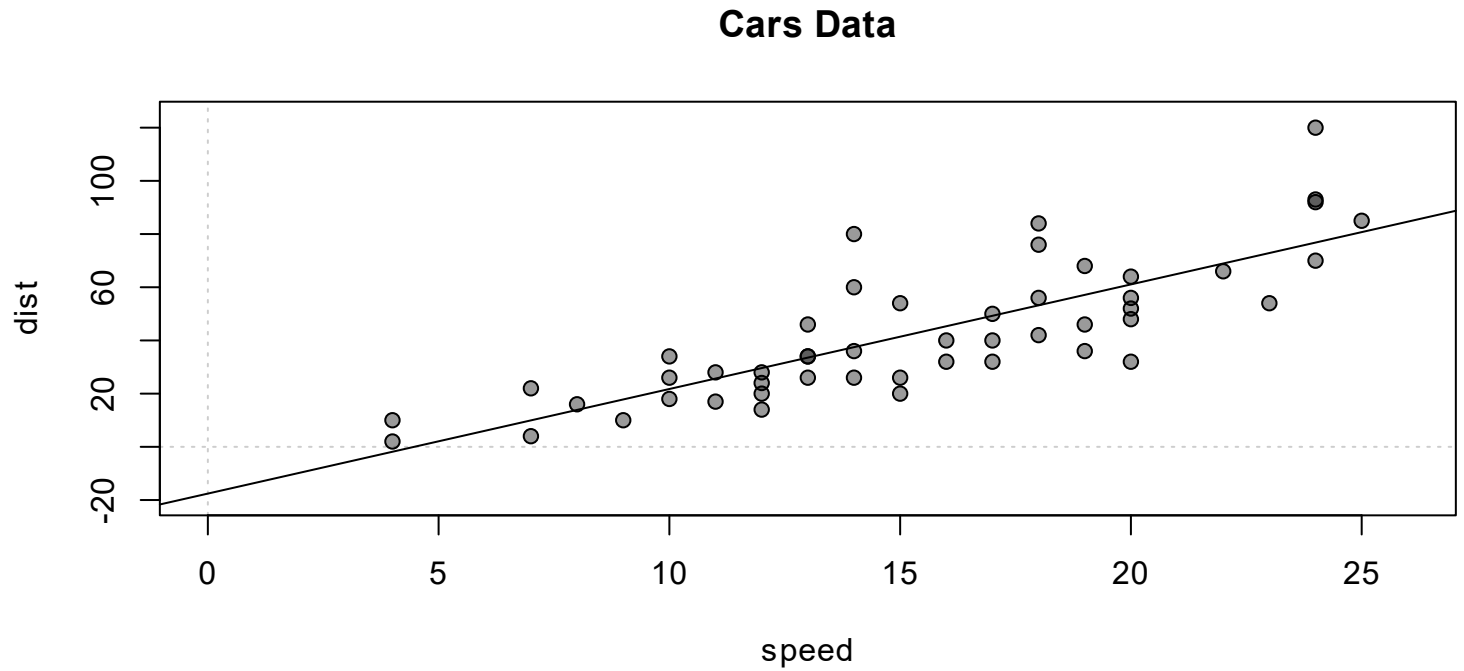
```
predict(fit_cars, newdata = data.frame(speed = 0))
        1
-17.57909
```

- Any potential issues with these predictions?
- Does that mean our model is *bad*?
- Is this value familiar to us from the model coefficient table?

Car Model: Building the Equation - Stopped Car

Take another look at a plot of the data and the model fit:

```
plot(dist ~ speed, data = cars, main = "Cars Data",
     xlim = c(0, 26), ylim = c(-20, 124))
abline(fit_cars)
```



0 speed is outside the range of our observations! It's an extrapolated value!

Simple and Multiple Linear Regressions

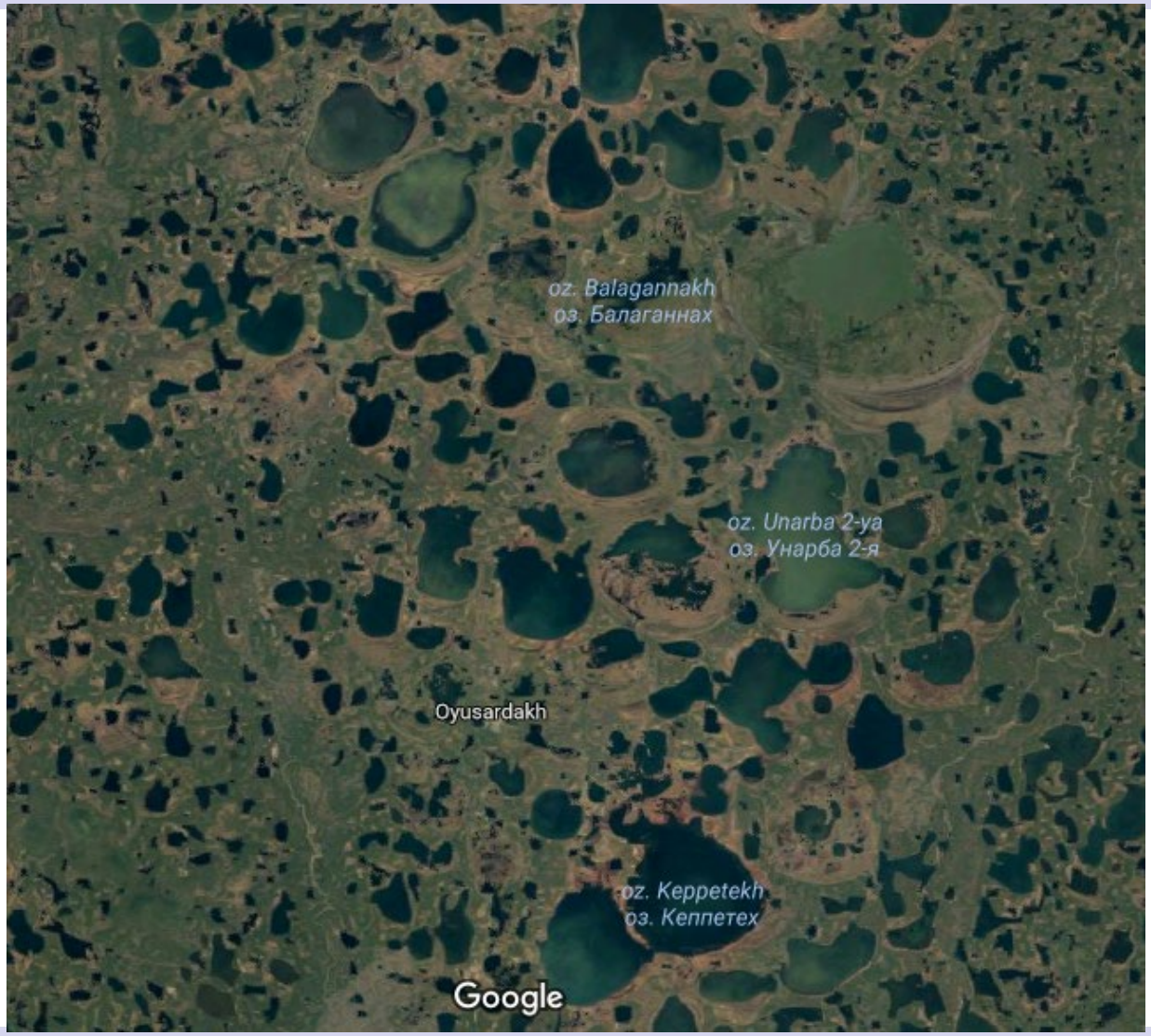
An Example

A more complex problem

In typical studies, we:	**we can use *multiple regression* when:**
• Rarely collect only a single explanatory variable • Are interested in *joint* effects • Are interested in *interactive* effects	• We have more than one explanatory variable. • Our explanatory variables are continuous. • We'll relax this requirement later.

An example – Pesticide Runoff and Fish Abundance

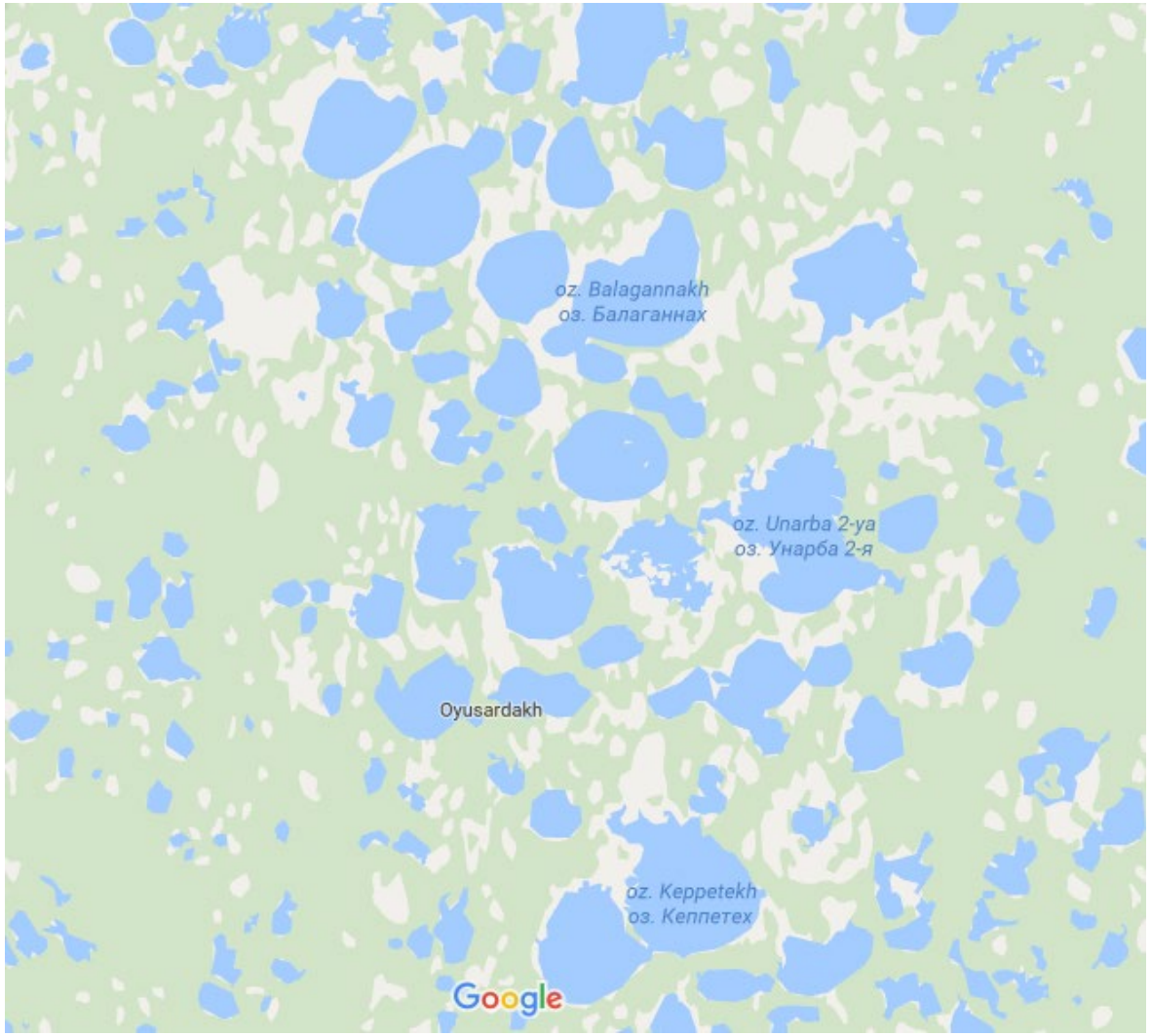
What might influence the number of a certain fish species (abundance) in each of these ponds?



An example – Pesticide Runoff and Fish Abundance

What might influence the number of a certain fish species (abundance) in each of these ponds?

- lake size: surface area
- pH
- connectivity
- depth
- human activity (e.g., fishing)
- agricultural run-off
- etc…



An example – Pesticide Runoff and Fish Abundance

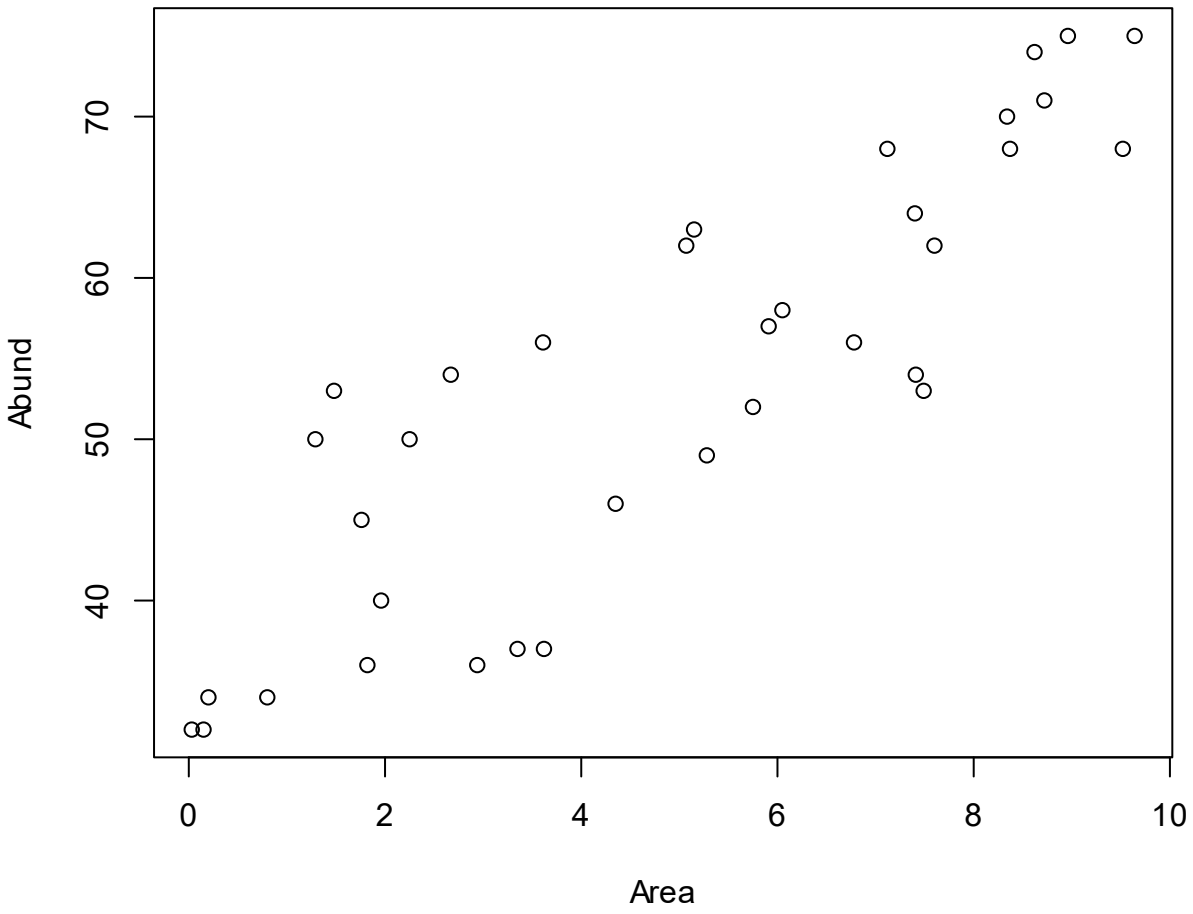
I am particularly interested in how:

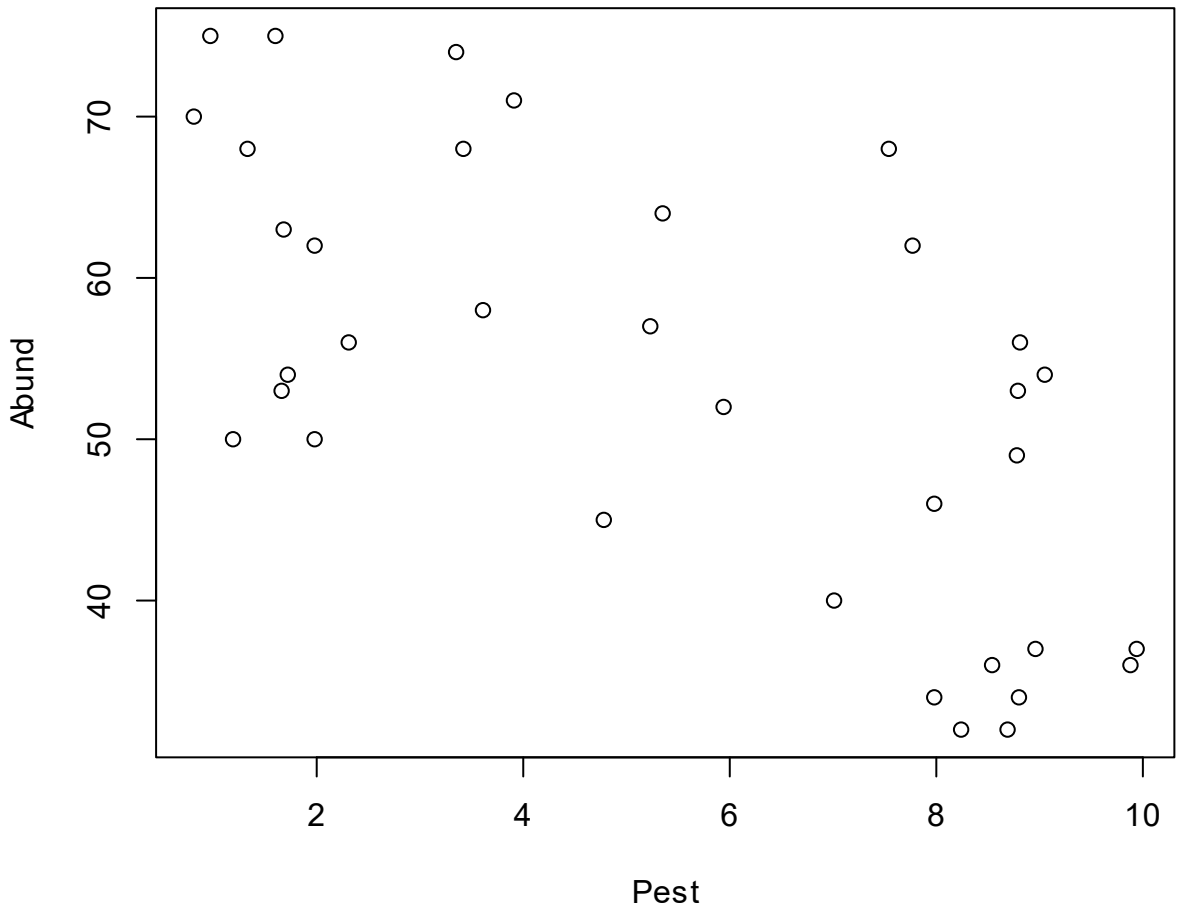
- the size of the lake (`Area`)
- the amount of pesticides in the water (`Pest`)

Might influence the number of fish counted in a lake (`Abundance`)

	Abundance	Area	Pest
1	34	0.80	7.98
2	68	9.52	7.54
3	71	8.72	3.91
4	68	7.12	3.42
5	58	6.05	3.61
6	50	2.25	1.98

Graphical Exploration: Scatterplots





Simple Regressions: Area and Pest

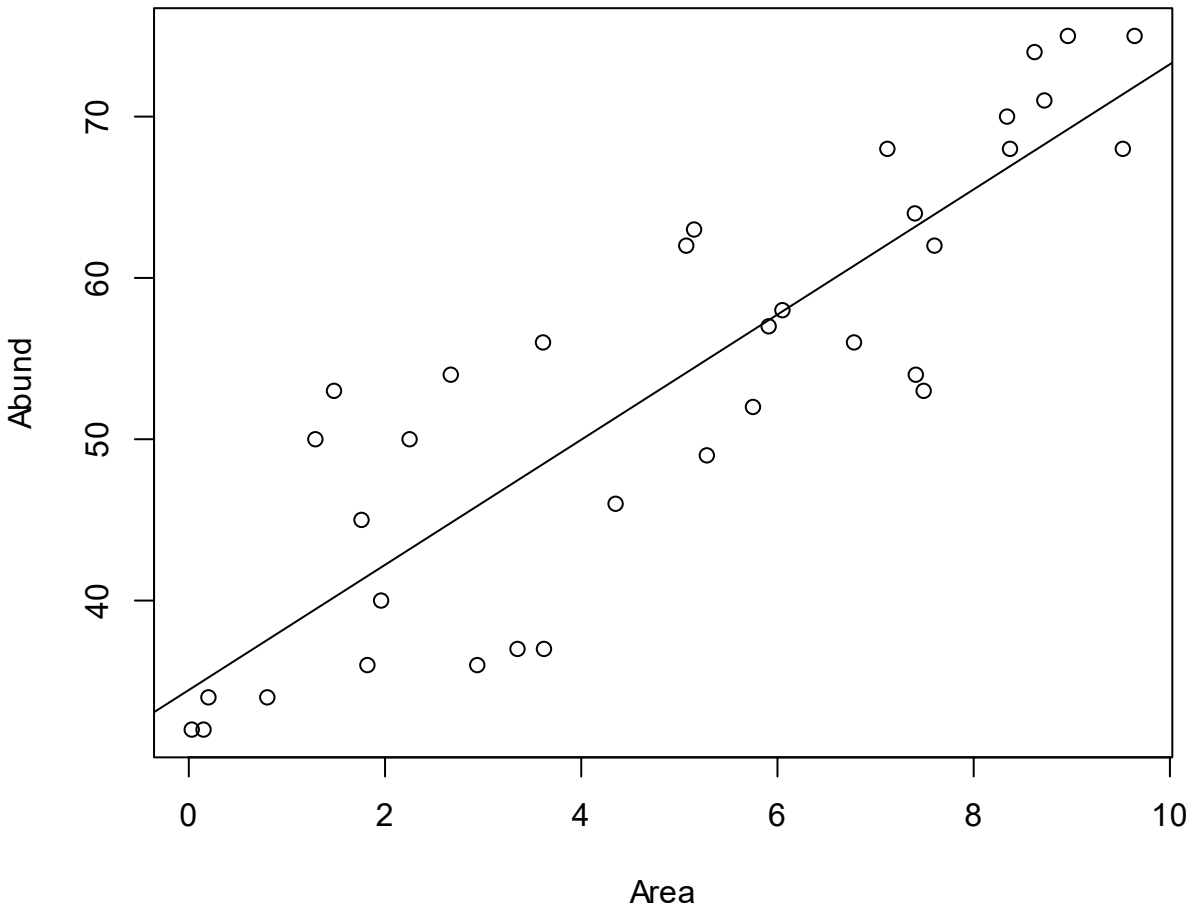
It looks like both factors are related to abundance.

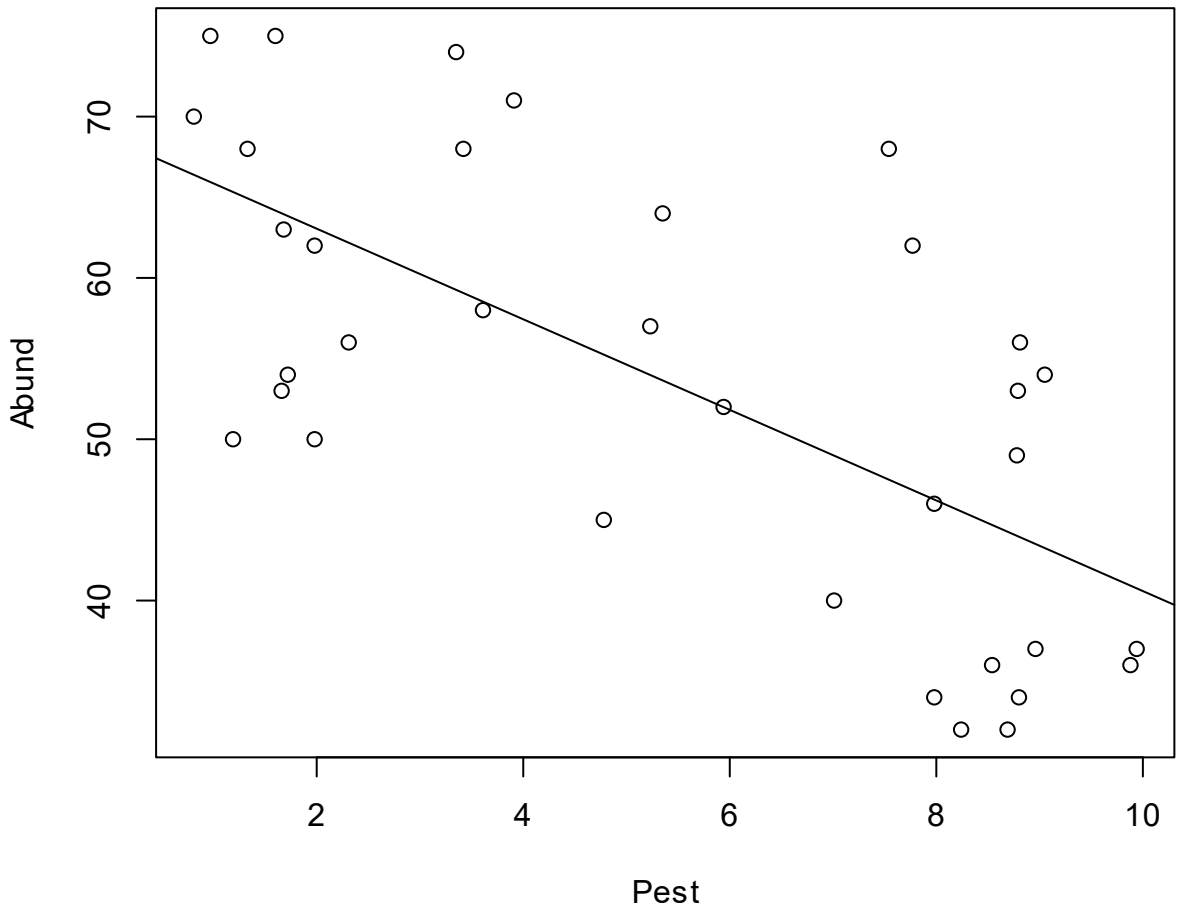
We'd like to know about both, how should we proceed?

We could make two models, one each for Area and Pest:

```
mod.area <- lm(Abundance ~ Area, data = dat_fish)

mod.pest <- lm(Abundance ~ Pest, data = dat_fish)
```

Model Plots





Both Models Look Good

- There are obvious relationships between each predictor and the response.
- How do the single-predictor models perform?

Area Model Summary

Positive slope

Model correlation = 0.75. The model explains 75% of the variation in the response

```
Coefficients:
             Estimate Std. Error t value Pr(>|t|)
(Intercept)  34.4537     2.1712   15.87  < 2e-16 ***
Area          3.8792     0.3787   10.24 8.82e-12 ***
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 6.674 on 33 degrees of freedom
Multiple R-squared:  0.7608,    Adjusted R-squared:  0.7535
F-statistic: 104.9 on 1 and 33 DF,  p-value: 8.819e-12
```

Pest Model Summary

Negative slope

Model correlation = 0.42. The model explains 42% of the variation in the response

```
Coefficients:
             Estimate Std. Error t value Pr(>|t|)
(Intercept)  68.6668     3.4289  20.026  < 2e-16 ***
Pest         -2.8080     0.5476  -5.128 1.26e-05 ***
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 10.18 on 33 degrees of freedom
Multiple R-squared:  0.4435,    Adjusted R-squared:  0.4266
F-statistic:  26.3 on 1 and 33 DF,  p-value: 1.265e-05
```

Simple Models Results

There is a significant positive relationship between Area and fish abundance

- $\beta_{\text{Area}} = 3.9$
- $p = 0$
- $R^2 = 0.75$

There is a significant negative relationship between Pests and fish abundance

- $\beta_{\text{Pest}} = -2.8$
- $p = 1.19e\text{-}5$
- $R^2 = 0.43$

Multiple Regression

Can we do better?

- It seems like both predictors are important, and we'd like to capture the effects of both in a single model…

Multiple Regression Model Table

Slope coefficients are slightly different than single-factor models, but both are highly significant.

Model explains 96% of the variation!

```
Coefficients:
             Estimate Std. Error t value Pr(>|t|)
(Intercept)  47.8594     1.2722   37.62  < 2e-16 ***
Pest         -1.9797     0.1423  -13.91 4.06e-15 ***
Area          3.3316     0.1501   22.20  < 2e-16 ***
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 2.553 on 32 degrees of freedom
Multiple R-squared:  0.9661,    Adjusted R-squared:  0.9639
F-statistic: 455.3 on 2 and 32 DF,  p-value: < 2.2e-16
```

Regression results: Multiple Regression

We found significant *joint* effects of `Area` & `Pest` in the multiple regression model:

- $\beta_{\text{Area}} = 3.33$ ($p = 0$)
- $\beta_{\text{Pest}} =$ -1.98 ($p = 0$)
- $R^2 = 0.96$

Correlated Predictors: Collinearity

Correlated Predictors

Whitebark Pine: Background

Richard Sniezko, US Forest Service - Forest Service Dorena lab

The whitebark pine, *Pinus albicaulis* is a high-altitude tree that grows in montane habitats in Western North America.

Whitebark Pine: Modeling

Warmer winters in recent decades are associated with many plant and animal species shifting their ranges to higher altitudes.

The seeds of whitebark pine is an important food source for many animals, including black bears.

Plant pests, including the Mountain Pine Beetle *Dendroctonus ponderosae* are also shifting their ranges, making novel (and susceptible) hosts available.

There is great interest in understanding, and predicting, how whitebark pine growth varies with altitude and temperature.

We could probably learn a lot by creating a regression model of pine growth predicted by average winter temperatures and altitude!

Whitebark Pine: Data

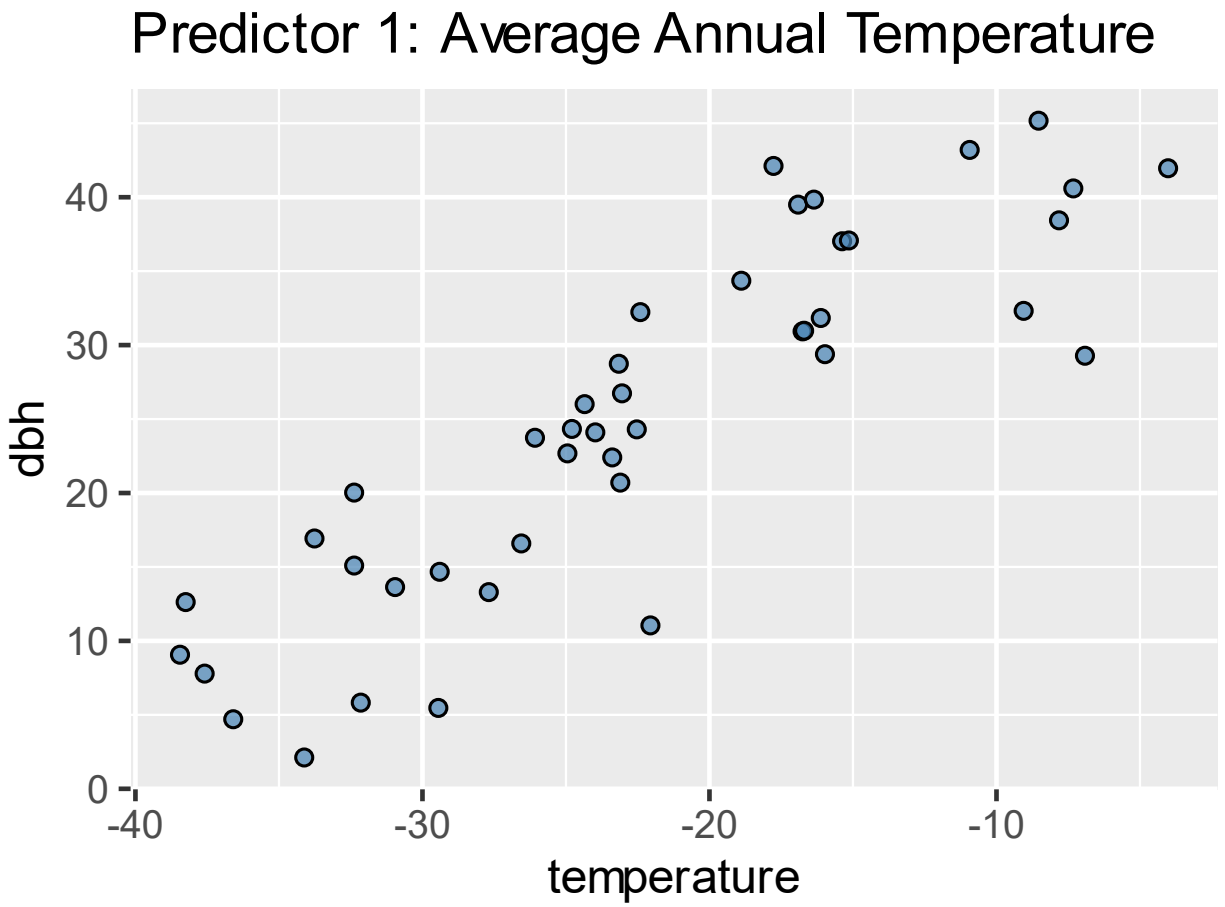
Suppose you have a dataset containing information about the size of individual trees `dbh`.

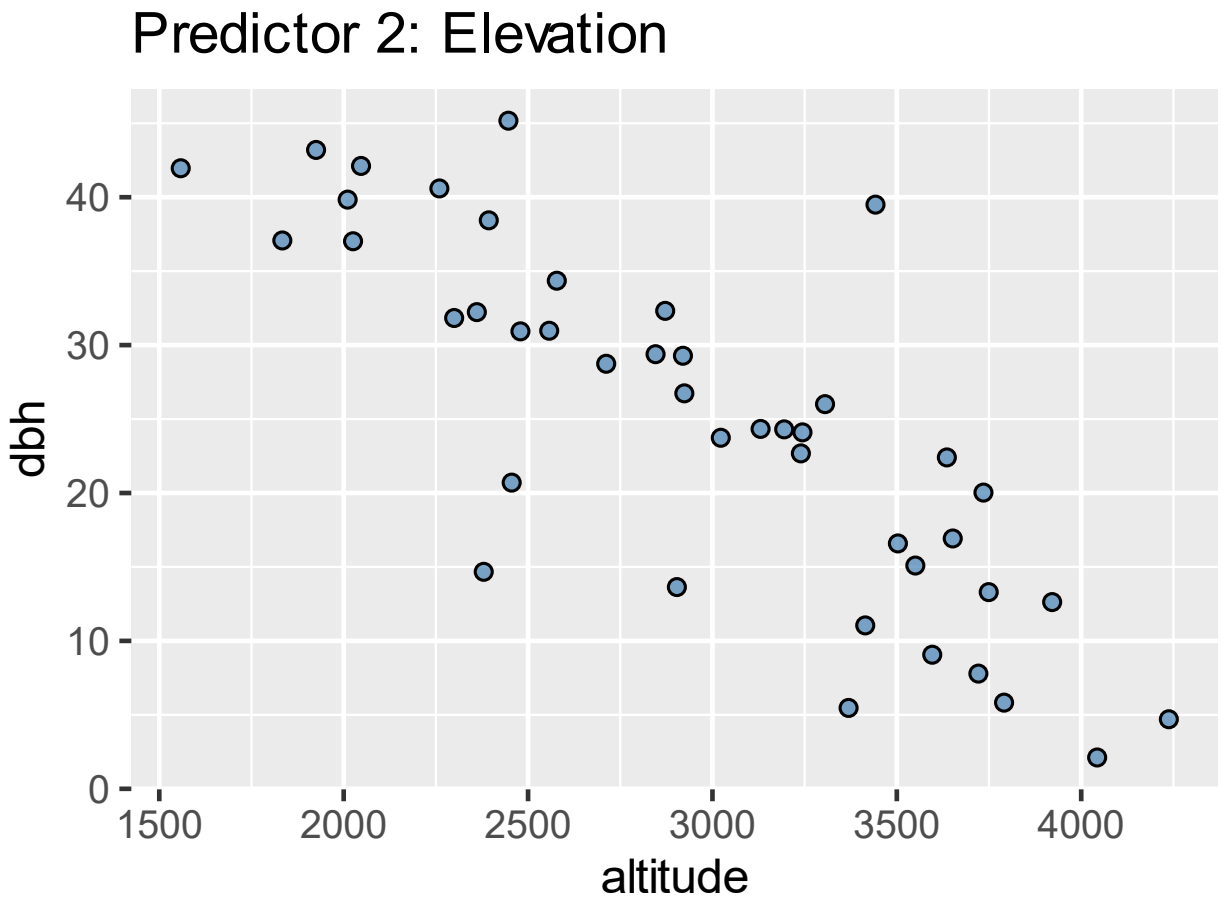
For each tree, you also know the altitude at which it grows, the man annual precipitation, and the mean annual temperature:

Do you think altitude and temperature are related?

```
head(dat_whitebark)
  X rainfall altitude      dbh
1 1 20.44596 2493.433 47.08338
2 2 33.23555 2301.189 52.09458
3 3 31.15216 2328.082 51.69528
4 4 25.24166 2390.111 49.53393
5 5 29.35912 2227.444 51.96837
6 6 28.51941 2421.535 49.85675
```

Whitebark Pine: Scatterplots

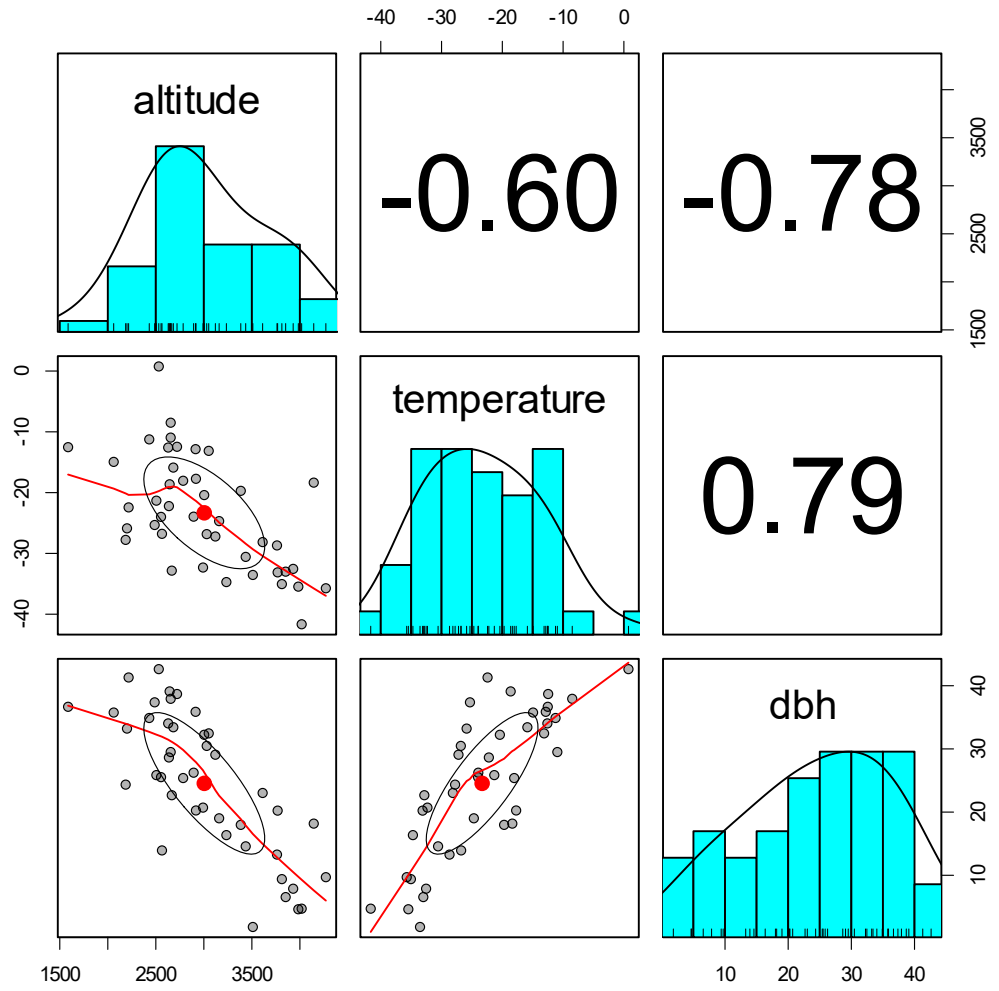




Whitebark Pine: Correlated Predictors

We know that they grow larger in warmer areas and at lower altitudes.

We can also see that altitude and temperature are strongly correlated in this **pair plot**:



Whitebark Pine: Simple Models

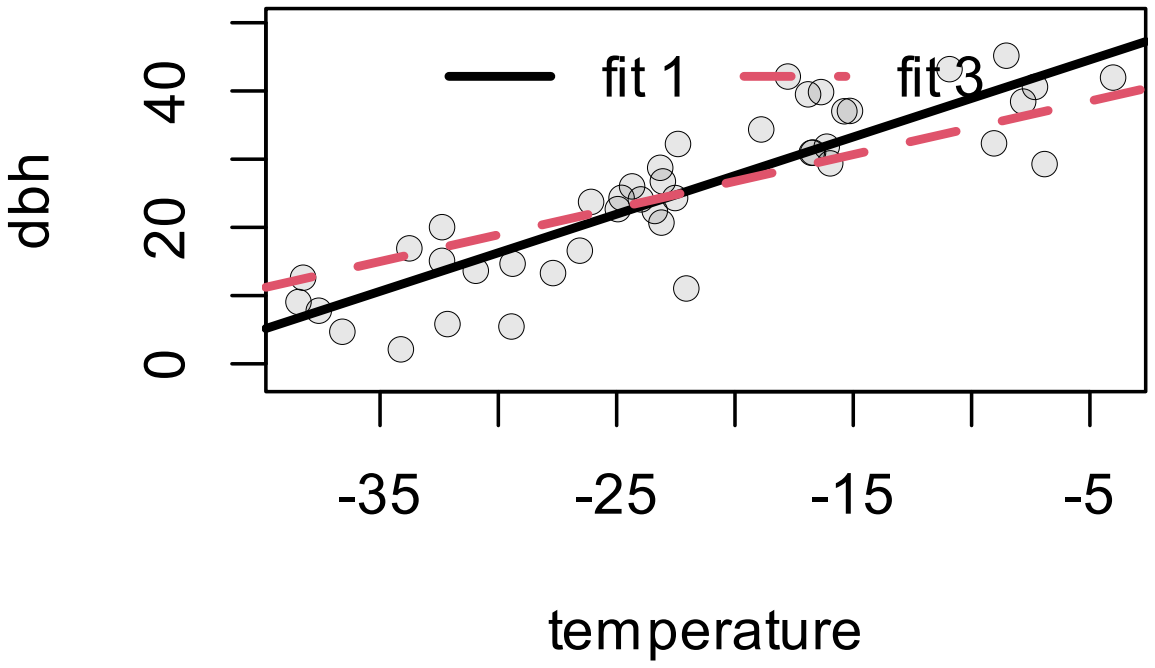
```
Call:
lm(formula = dbh ~ temperature, data =
dat_whitebark_collinear)

Residuals:
    Min      1Q  Median      3Q     Max
-14.014  -3.183   1.095   3.618  10.786

Coefficients:
             Estimate Std. Error t value Pr(>|t|)
(Intercept) 47.08033    2.05861   22.87  < 2e-16 ***
temperature  1.02641    0.08136   12.62 1.06e-15 ***
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1
' ' 1

Residual standard error: 5.881 on 41 degrees of freedom
Multiple R-squared:  0.7952,    Adjusted R-squared:
0.7902
F-statistic: 159.2 on 1 and 41 DF,  p-value: 1.057e-15
```

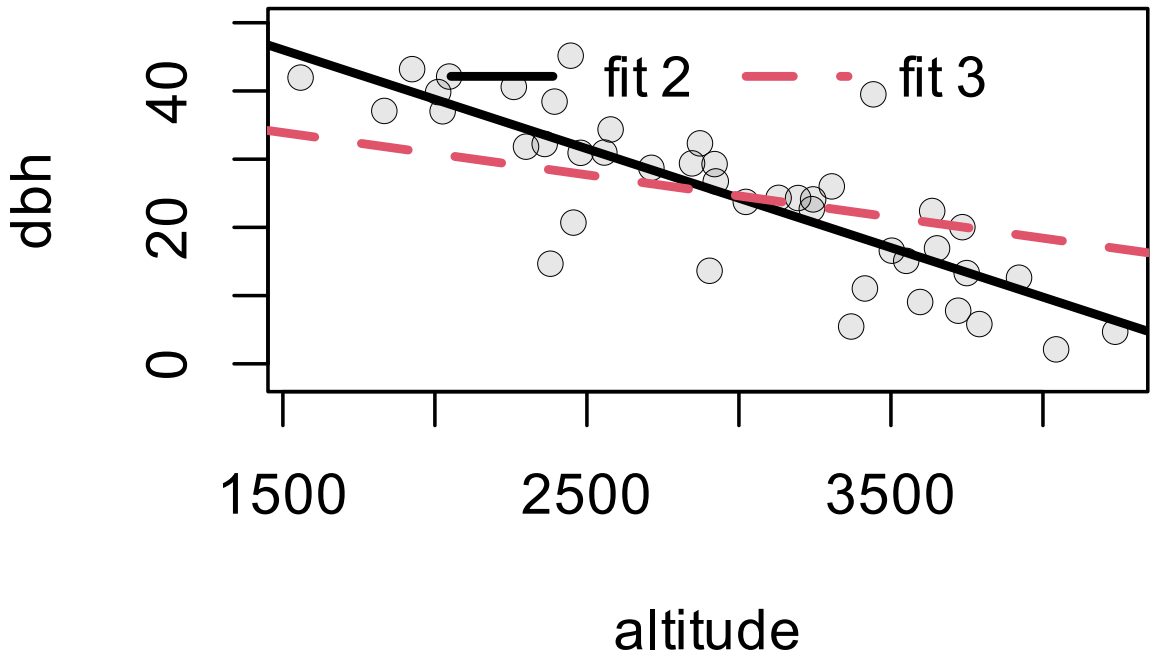
```
Call:
lm(formula = dbh ~ altitude, data =
dat_whitebark_collinear)

Residuals:
    Min      1Q  Median      3Q     Max
-13.457  -5.829  -2.056   3.863  18.670

Coefficients:
             Estimate Std. Error t value Pr(>|t|)
(Intercept) 70.883870   6.053708  11.709 1.17e-14 ***
altitude    -0.015641   0.001964  -7.963 7.46e-10 ***
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1
' ' 1

Residual standard error: 8.143 on 41 degrees of freedom
Multiple R-squared:  0.6073,    Adjusted R-squared:
0.5977
F-statistic: 63.41 on 1 and 41 DF,  p-value: 7.462e-10
```

Correlated Predictors: Multiple Regression

```
Call:
lm(formula = dbh ~ altitude + temperature, data = dat_whitebark_collinear)

Residuals:
     Min       1Q   Median       3Q      Max
-14.1944  -3.6237   0.4966   3.9435   9.1340

Coefficients:
             Estimate Std. Error t value Pr(>|t|)
(Intercept) 58.918987   4.301265  13.698  < 2e-16 ***
altitude    -0.005705   0.001865  -3.059  0.00395 **
temperature  0.790577   0.106972   7.390 5.39e-09 ***
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 5.36 on 40 degrees of freedom
Multiple R-squared:  0.834, Adjusted R-squared:  0.8257
F-statistic: 100.5 on 2 and 40 DF,  p-value: 2.527e-16
```

Correlated Predictors: What happened?





- The coefficients changed. They are now obviously wrong!
- Significance changed: both are still significant, but p-values are higher.

This seems very weird.

Our 2-predictor model performed well for the fish/area/pesticide model.

Why did we have a problem?

Correlated Predictors: Collinearity

Collinearity: when two or more predictors are highly correlated with each other

- Highly correlated predictors contain the **same information**.
- Since they contain the same info, the model can't determine which variable to attribute the effect to!
- The mathematical reasoning is that correlated predictors cause the **design matrix** to be less than **full rank**.
- Don't worry if you don't know what this means, it's not essential for understanding the problem.

What to do?

- Examine a `pair plot`. Base R has the `pairs()` function. Package `psych` has a nice function called "pairs.panels()"
- Remove one of the highly correlated predictors.
- Check for variance inflation using `vif()`

Model Diagnostics 1: Normality of the Residuals

Model Diagnostics: Assumptions

Remember those pesky assumptions that we made?

Two of the most important were:

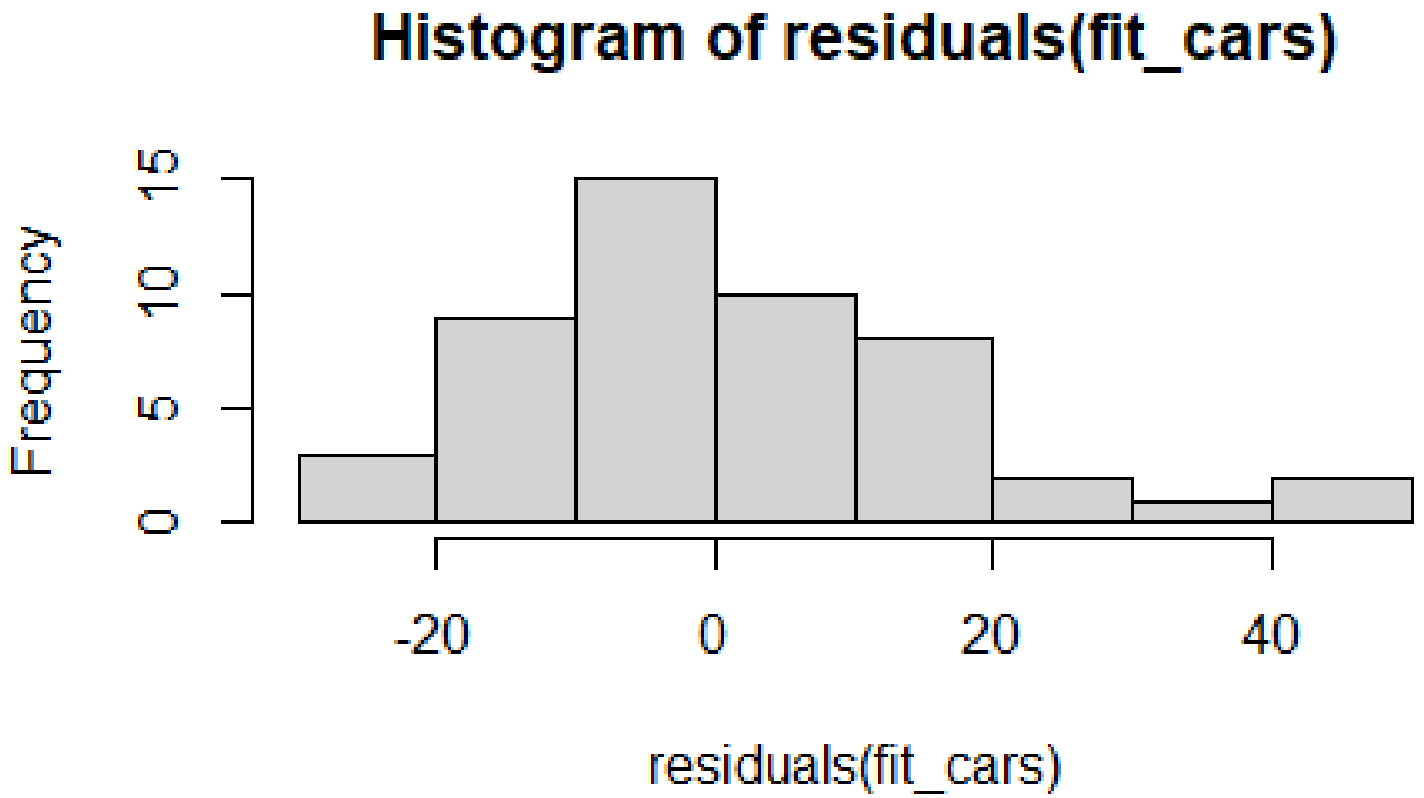
- Model residuals are Normally distributed
- Observations are independent

Model Diagnostics: Normality of Residuals

A histogram provides a quick visual check:

hist(residuals(fit_cars))

The two extreme values in the rightmost bin may be troublesome



Model Diagnostics: Normality of Residuals

The Shapiro tests formalizes what we see in the histogram:

```
shapiro.test(residuals(fit_cars))

    Shapiro-Wilk normality test

data:  residuals(fit_cars)
W = 0.94509, p-value = 0.02152
```

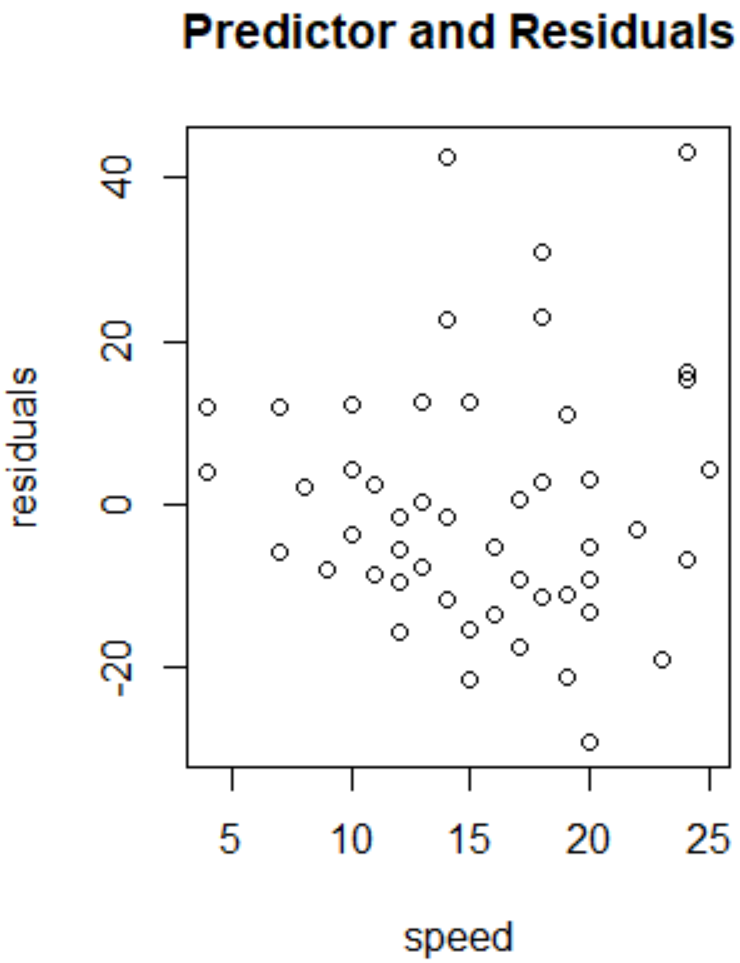
There is evidence of non-normality, but it may be due to just a few extreme values.

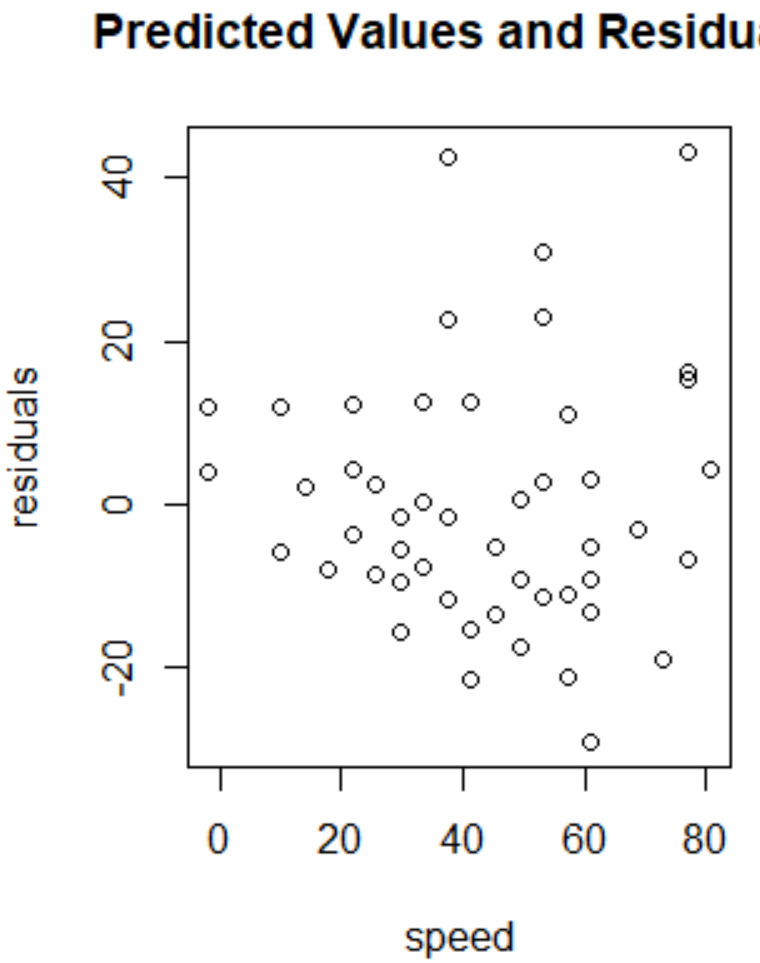
Model Diagnostics: Independence of Residuals

Scatterplots of the residuals against the predictors and the predicted values can help detect non-independence or other potential issues:

```
par(mfrow = c(1, 2), oma = c(0, 0, 0,
0))
plot(
  x = cars$speed, y =
residuals(fit_cars),
  main = "Predictor and Residuals",
  xlab = "speed", ylab = "residuals")
plot(
  x = predict(fit_cars), y =
residuals(fit_cars),
  main = "Predicted Values and
Residuals",
  xlab = "speed", ylab = "residuals")
```

Model Diagnostics: Independence of Residuals Plots





Model Diagnostics

In the scatterplot visual assessments, you are looking for any obvious patterns in the data.

Note that it can be hard to detect subtle problems visually.

The visual and numerical diagnostics suggest there might be some issues with non-normality.