University of East Anglia

BIO-5023YB
2020
Spring term – week 6
Multiple Predictors & Interaction effects

Dr Philip Leftwich – p.leftwich@uea.ac.uk

Learning outcomes

- Multiple predictors

- Interaction terms

- Results Bias

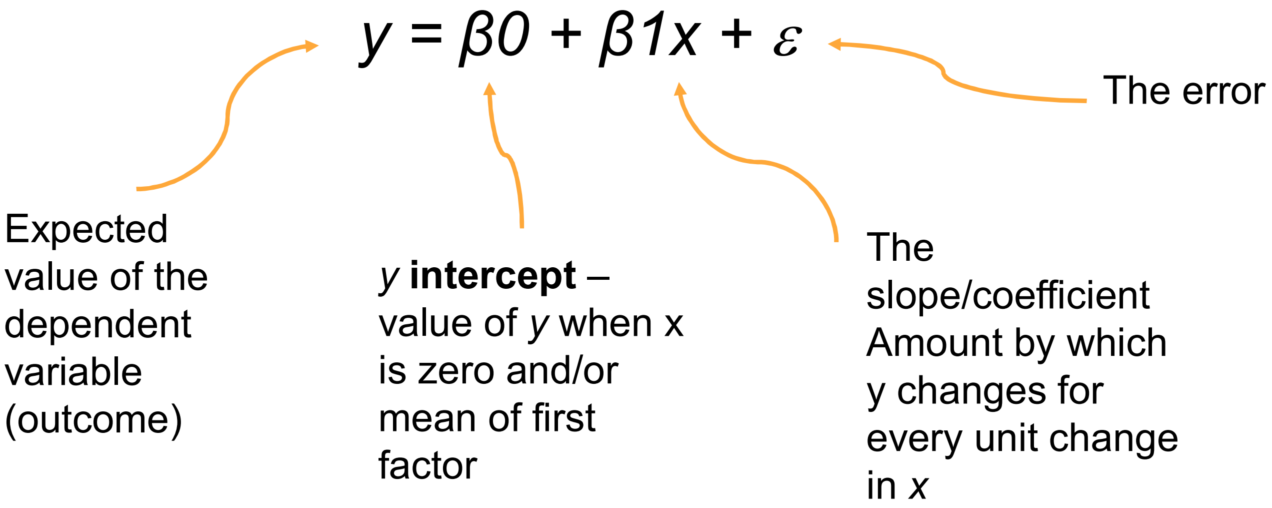
Introduction

In previous weeks we have learned about linear regression by ordinary least squares

$$y = \beta 0 + \beta 1 * x$$

Introduction

In previous weeks we have learned about linear regression by ordinary least squares



$$y = \beta 0 + \beta 1x + \varepsilon$$

The error

Expected value of the dependent variable (outcome)

y **intercept** – value of *y* when x is zero and/or mean of first factor

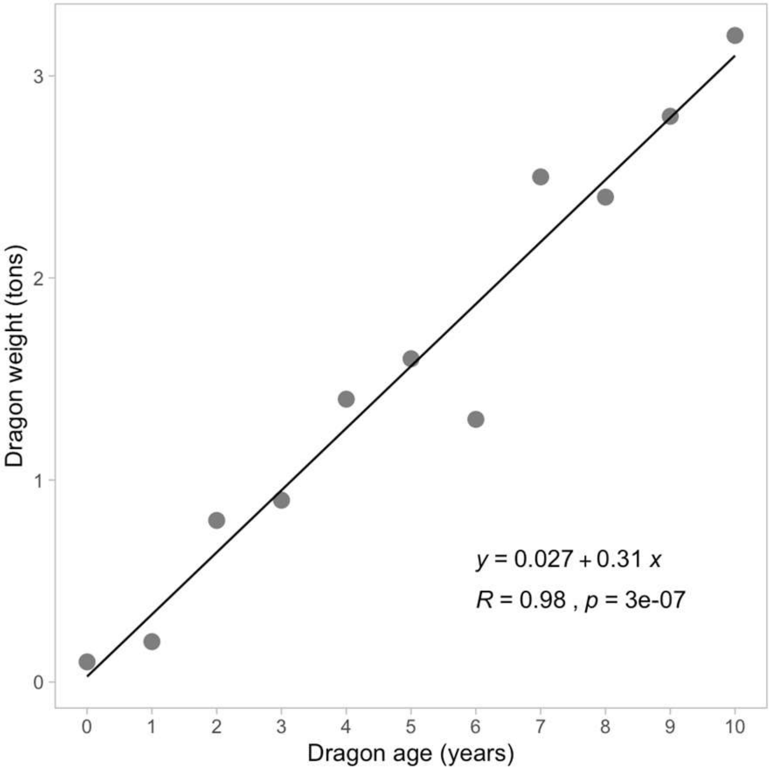
The slope/coefficient Amount by which y changes for every unit change in *x*

Introduction

At age zero/birth we expect dragon weight to be_______

For each 1 year increase in dragon age we expect weight to increase by an *average* of ______ tons

At 4 years old we expect an average weight of a dragon to be __________ ?

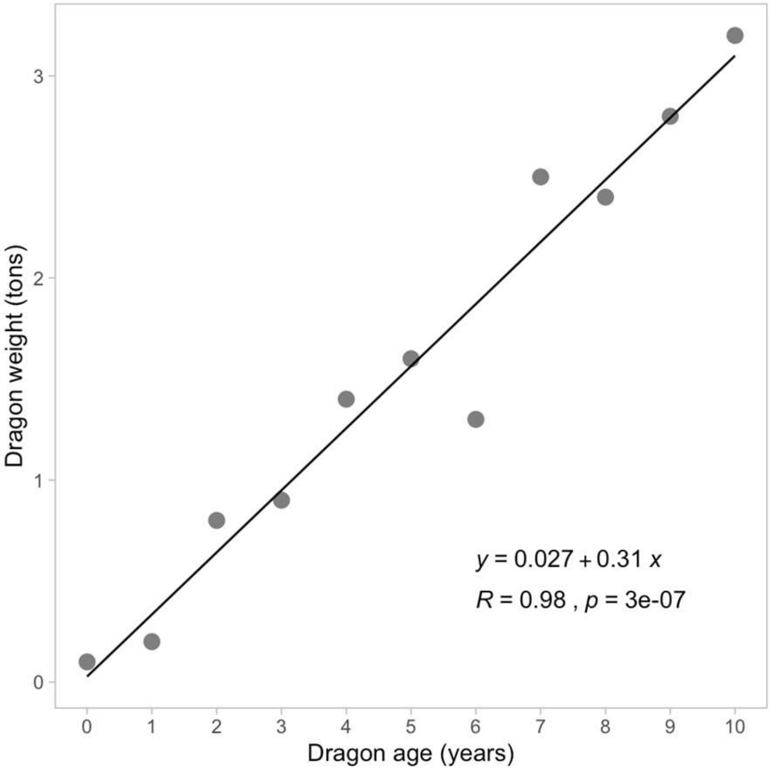


Making predictions

At age zero/birth we expect dragon weight to be **0.027 tons**

For each 1 year increase in dragon age we expect weight to increase by an *average* of **0.31 tons**

At 4 years old we expect an average weight of a dragon to be **1.267 tons**



Single predictor, single outcome

In that example, we could consider dragon age as a single predictor variable, and dragon weight as a single continuous outcome variable.

But often, we don't just have a single predictor variable that influences a single continuous outcome variable - we can imagine that the outcome variable **dragon weight** might be influenced by a number of variables, like: age, species, and diet.

When we are trying to explore relationships between **multiple predictor variables** (continuous or categorical) and a **single continuous outcome variable**, we might use *multiple linear regression.*

Multiple linear regression

Now let’s consider one continuous **outcome** variable (**dependent**) and **two** continuous independent **predictor** variables.

```
dragon_mlr <- lm(dragon_wt ~ dragon_age + dragon_diet, data = dragon_df_2)
```

Coefficients:

Intercept
0.061444

Dragon_age
0.308176

Dragon_diet
-0.001235

dragon_age	dragon_diet	dragon_wt
0	20.2	0.1
1	30.6	0.2
2	40.2	0.8
3	31.6	0.9
4	19.2	1.4
5	45.6	1.6
6	26.1	1.3
7	20.4	2.5
8	31.8	2.4
9	50.4	2.8

Multiple predictors

Multiple linear regression

Written as an equation

```
dragon_mlr <- lm(dragon_wt ~ dragon_age + dragon_diet, data = dragon_df_2)
```

Coefficients:

Intercept
0.061444

Dragon_age
0.308176

Dragon_diet
-0.001235

weight(age, diet) = 0.0614 + 0.3082*(age) -0.0012*(diet)

Multiple linear regression

weight(age, diet) = 0.0614 + 0.3082*(age) -0.0012*(diet)

Based on our model what weight would we expect for a dragon that is 8.5 years old and eats 45 pounds of food per day

Example: weight(8.5, 45) = 0.0614 + 0.3082*(**8.5**) -0.0012*(**45**) = **2.6254 tons**

Based on our model, what weight would we expect for a dragon that is 2.2 years old and eats 18.3 pounds of food per day?

Example: weight(2.2, 18.3) = 0.0614 + 0.3082*(**2.2**) -0.0012*(**18.3**) = **0.7168 tons**

Interpret coefficients

How do we **interpret** the individual coefficients associated with each of the predictor variables?

Coefficients for continuous predictor variables in multiple linear regression can be interpreted as

*the expected change in value of the outcome variable with each 1-unit increase in the predictor variable, **if everything else is held constant**.*

> **For each one year increase in dragon age we expect dragon weight to increase by 0.3082 tons on average, if dragon diet is held constant.**
>
> Another way to think about this: If we have two dragons that eat the same daily amount (diet is constant) but are separated in age by one year, then we'd expect the older dragon to weigh 0.3082 tons more than the younder dragon, on average.
>
> **For each one pound increase in dragon daily diet we expect dragon weight to decrease by 0.0012 tons, if dragon age is held constant.**
>
> Another way to think about this: If we have two dragons that are the same exact age, if Dragon A eats one pound of food more daily than Dragon B, we'd expect Dragon A to weigh 0.0012 tons LESS than Dragon B, on average

Coefficients for categorical predictors

We are already familiar with *factors* for categories

If we have **three levels** to a factor – by default R will code the levels in alphabetical order so that:

For 3 species of dragon:

- FriendlyGreen
- GiantPurple
- MiniBlue

FriendlyGreen will be coded as the reference level and set as the Intercept value

Coefficients for categorical predictors

We are already familiar with *factors* for categories

$$\mathbf{weight(age,\ diet) = 0.0614 + 0.3082*(age)\ -0.0012*(diet)}$$

$$\mathbf{weight(age,\ diet,species) = 0.41 + 0.283*(age)\ +0.019*(diet)\ -\ 0.056*(MiniBlue) + 0.077*(GiantPurple)}$$

Why have the estimates for age and diet changed?

Why isn’t FriendlyGreen in the equation?

Coefficients for categorical predictors

We are already familiar with *factors* for categories

$$\mathbf{weight(age,\ diet) = 0.0614 + 0.3082*(age) -0.0012*(diet)}$$

$$\mathbf{weight(age,\ diet,species) = 0.41 + 0.283*(age) +0.019*(diet) - 0.056*(MiniBlue) + 0.077*(GiantPurple)}$$

Why have the estimates for age and diet changed? – **Because these are now the estimates for a FriendlyGreen**

Why isn’t FriendlyGreen in the equation? – **FriendlyGreen is our species reference the other values represent deviation from these values by species**

Coefficients for categorical predictors

weight(age, diet,species) = 0.41 + 0.283*(age) +0.019*(diet) - 0.056*(MiniBlue) + 0.077*(GiantPurple)

0.283*age - For each one year increase in dragon age we expect dragon weight to increase by 0.283 tons on average, if dragon diet is held constant.

0.019*diet - For each one pound increase in dragon daily diet we expect dragon weight to increase by 0.019 tons, if dragon age is held constant.

-0.056*(MiniBlue) – If everything about the dragons is the same (age,diet) then we expect a MiniBlue to weigh less than a FriendlyGreen by -0.056 tons on average

0.019*(GiantPurple) - If everything about the dragons is the same (age,diet) then we expect a GiantPurple to weigh more than a FriendlyGreen by 0.019 tons on average

Coefficients for categorical predictors

weight(age, diet,species) = 0.41 + 0.283*(age) +0.019*(diet) - 0.056*(MiniBlue) + 0.077*(GiantPurple)

Based on our model what weight would we expect for a MiniBlue dragon that is 5.7 years old and eats 25.1 pounds of food per day

Example: weight(5.7, 25.1, MiniBlue) = 0.41 + 0.283*(**5.7**) + 0.019*(**25.1**) -0.056*(**1**) +0.077*(0) =

2.44 tons

Note Categorical Predictors are mutually exclusive – a dragon can be *either* FriendlyGreen, MiniBlue or GiantPurple

Multiple linear regressions

- **These are probably the most used types of regression you will come across**

- **Our standard assumptions of linear models must be considered**
- Linear Relationship – Normality of Residuals – Homoscedasticity
- Independence of Residuals

Practice questions

There are some practice questions for you to try on this week's worksheet. Complete and check them to test your understanding

Interactions

Interaction terms in Multiple Linear Regression

Why?

- You think that one of the predictor variables *changes the WAY that ANOTHER predictor variable influences the* outcome variable

- Increase understanding/exploration of predictor variable effects

- Test for an interaction between predictor variables

Why not?

- Complicates interpretation of coefficients

- Increased model complexity

Example:

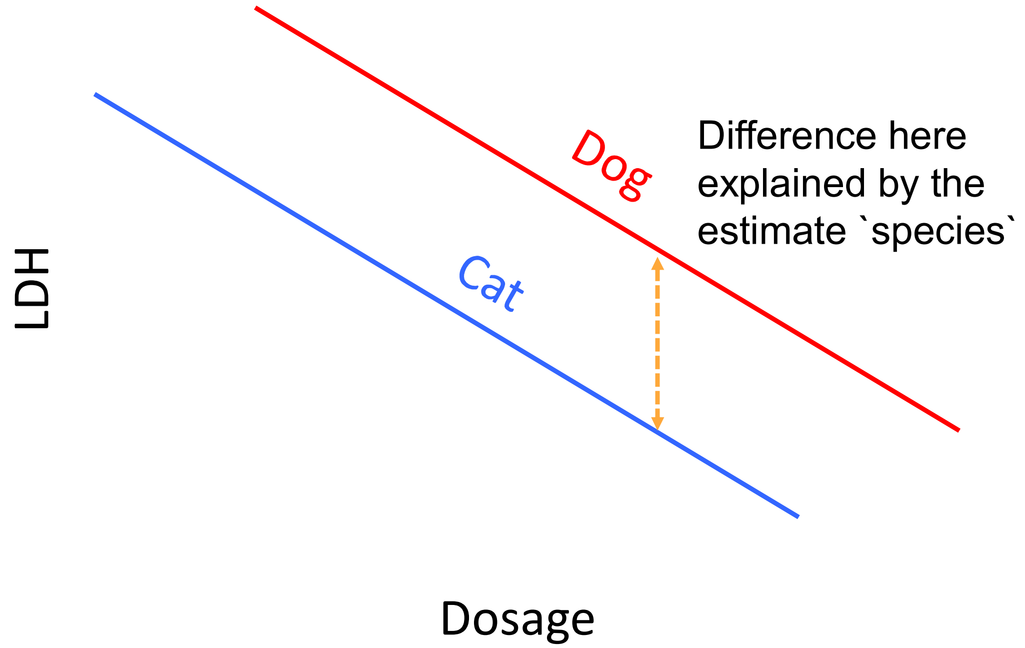
You are investigating the effects of drug dosage on LDH levels in cats and dogs (using Species and Dosage as the predictor variables).

You perform multiple linear regression, which looks something like this:

$$LDH = \beta_0 + \beta_1 * Species + \beta_2 * Dose$$

LDH – Lactate dehydrogenase – used as an indicator of tissue damage

If there was NO interactive effect of Species on DOSAGE (i.e., we do not expected that Species will change the way DOSAGE influences LDH levels), we could expect something like this:

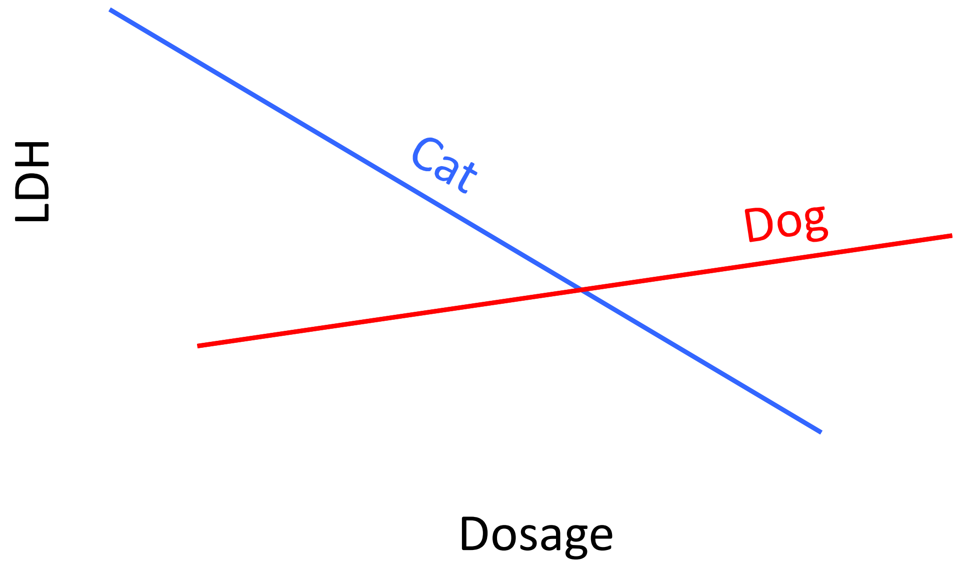


Difference here explained by the estimate `species`

Notice: Species doesn't *change the WAY that* dosage influences LDH levels. There is NO interactive effect between species and dosage.

$LDH = \beta_0 + \beta_1 * Species + \beta_2 * Dose$

If there IS AN interactive effect between Species and Dosage (i.e., Species will change the *way* DOSAGE influences LDH levels), we could expect something like this:



Notice: Species CHANGES *the WAY that* dosage influences LDH levels. There *is* an interactive effect between species and dosage.

If you suspect an interactive effect and/or want to test for it, you might consider adding an interaction term to your regression model:

$$\text{LDH} = \beta_0 + \beta_1 {*}\text{Species} + \beta_2 {*}\text{Dose} + \beta_3 {*}(\underbrace{\text{Species}{*}\text{Dose}}_{\text{Interaction Term}})$$

Interpreting interaction terms

A researcher uses multiple linear regression to model algae concentration (mg/L) as a function of water temperature (Temp, ℃) and phosphate concentration (mg/L).

First, let’s consider a model without an interaction term:

Algae = 4.2 + 0.5*Temp + 1.2*Phosphate

How do we interpret these coefficients?

But what if we add an interaction term?

Algae = 4.2 + 0.7*Temp + 1.2*Phosphate + 0.4*Temp*Phosphate

Then we can’t just interpret **Temp** as “on average, algae concentration is expected to increase by 0.7 mg/L for each 1 degree increase in temperature” because **Temp** *also shows up in the interactive term.*

We’d have to interpret it, for example, at a specific phosphate concentration of 10 mg/L: “If phosphate is 10 mg/L, we expect algae concentration to increase by (0.7 + 0.4*10) mg/L for each degree increase in temperature.”

So, when it comes to interaction terms:

- <u>Have a conceptual basis for why you'd expect an interaction</u> between Explanatory Variables (and visualize it)

- Know that your coefficients will CHANGE when you add an interaction term, and you HAVE to consider the interaction when you are interpreting the coefficients

- You can add the interaction term as a way to TEST the hypothesis that there is no significant interaction between Explanatory Variables and THEN remove it

Results Bias

Bias: a systematic difference between an estimator (e.g. value, coefficient, etc.) and a true population parameter

(1) Selection bias: Bias created by the non-random collection of observations – Planning stage and/or data collection stage

(2) Omitted-variable bias (OVB): Bias created when one or more predictor variables are incorrectly left out of a model – Analysis stage

Selection Bias:

- Sampling bias – Non-random collection of observations
- Data bias – Arbitrary and/or incorrect group of or omission of observations
- Study bias – Reporting only ‘favorable’ results
- Time interval bias – Ending (or beginning) trials or observations at times that might non-randomly influence the outcome

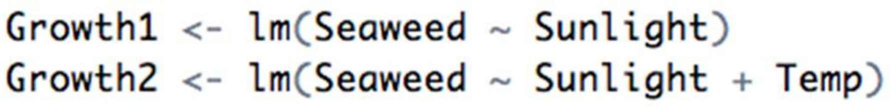
Omitted Variable Bias (OVB):

Bias created when one or more predictor variables are incorrectly left out of a model

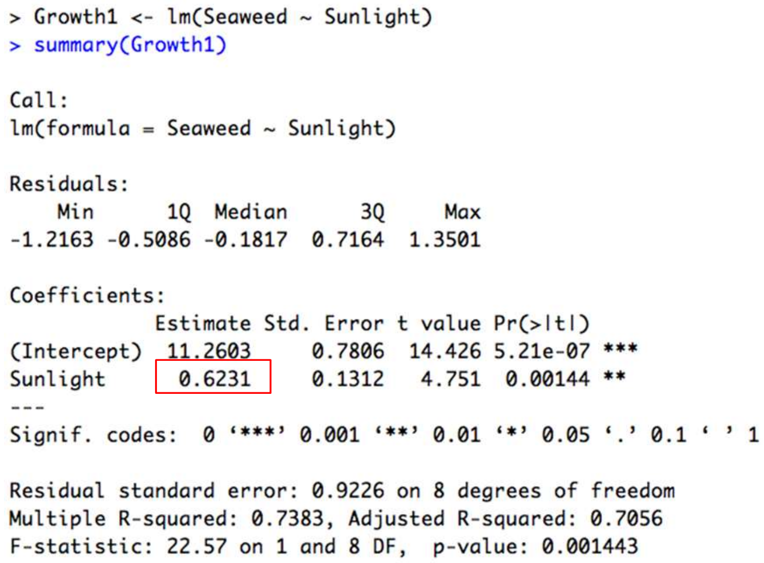
…i.e., it *should* be included as a predictor variable because you expect it to influence the outcome.

OVB Effects:

- Other variable coefficients are incorrectly over- or underestimated (and it’s not always clear which will occur)
- Standard errors and variances are wrong, and there’s no way to know which direction they’re wrong in
- Significance tests & CIs biased and can lead to wrong decisions
- Predictions based on model will be wrong (biased)



```
Growth1 <- lm(Seaweed ~ Sunlight)
Growth2 <- lm(Seaweed ~ Sunlight + Temp)
```



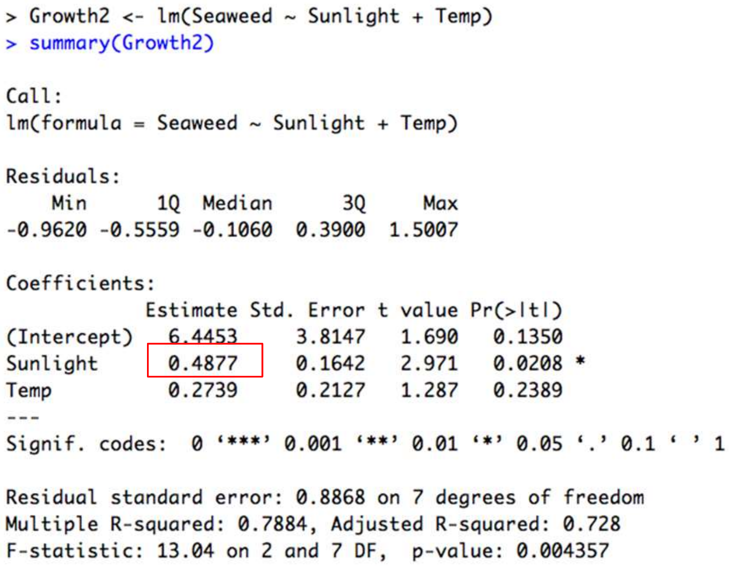
```
> Growth1 <- lm(Seaweed ~ Sunlight)
> summary(Growth1)

Call:
lm(formula = Seaweed ~ Sunlight)

Residuals:
    Min      1Q  Median      3Q     Max 
-1.2163 -0.5086 -0.1817  0.7164  1.3501 

Coefficients:
            Estimate Std. Error t value Pr(>|t|)    
(Intercept)  11.2603     0.7806  14.426 5.21e-07 ***
Sunlight      0.6231     0.1312   4.751  0.00144 ** 
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 0.9226 on 8 degrees of freedom
Multiple R-squared: 0.7383, Adjusted R-squared: 0.7056 
F-statistic: 22.57 on 1 and 8 DF,  p-value: 0.001443
```



```
> Growth2 <- lm(Seaweed ~ Sunlight + Temp)
> summary(Growth2)

Call:
lm(formula = Seaweed ~ Sunlight + Temp)

Residuals:
    Min      1Q  Median      3Q     Max 
-0.9620 -0.5559 -0.1060  0.3900  1.5007 

Coefficients:
            Estimate Std. Error t value Pr(>|t|)  
(Intercept)   6.4453     3.8147   1.690   0.1350  
Sunlight      0.4877     0.1642   2.971   0.0208 *
Temp          0.2739     0.2127   1.287   0.2389  
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 0.8868 on 7 degrees of freedom
Multiple R-squared: 0.7884, Adjusted R-squared: 0.728 
F-statistic: 13.04 on 2 and 7 DF,  p-value: 0.004357
```

In the absence of “temperature,” the influence of sunlight on kelp growth is overestimated

A special case:

Simpson’s Paradox: A similar trend occurring in different groups (levels) can disappear or reverse when the groups are combined (i.e., when one of the variables is removed).

Simpson’s Paradox Example:

You are trying to understand the relationship between bill length and bill depth in the Palmer Penguins dataset.

Imagine two hypotheses

1) Bill length and depth are correlated

2) Bill length and depth are correlated, but we need to take species into consideration

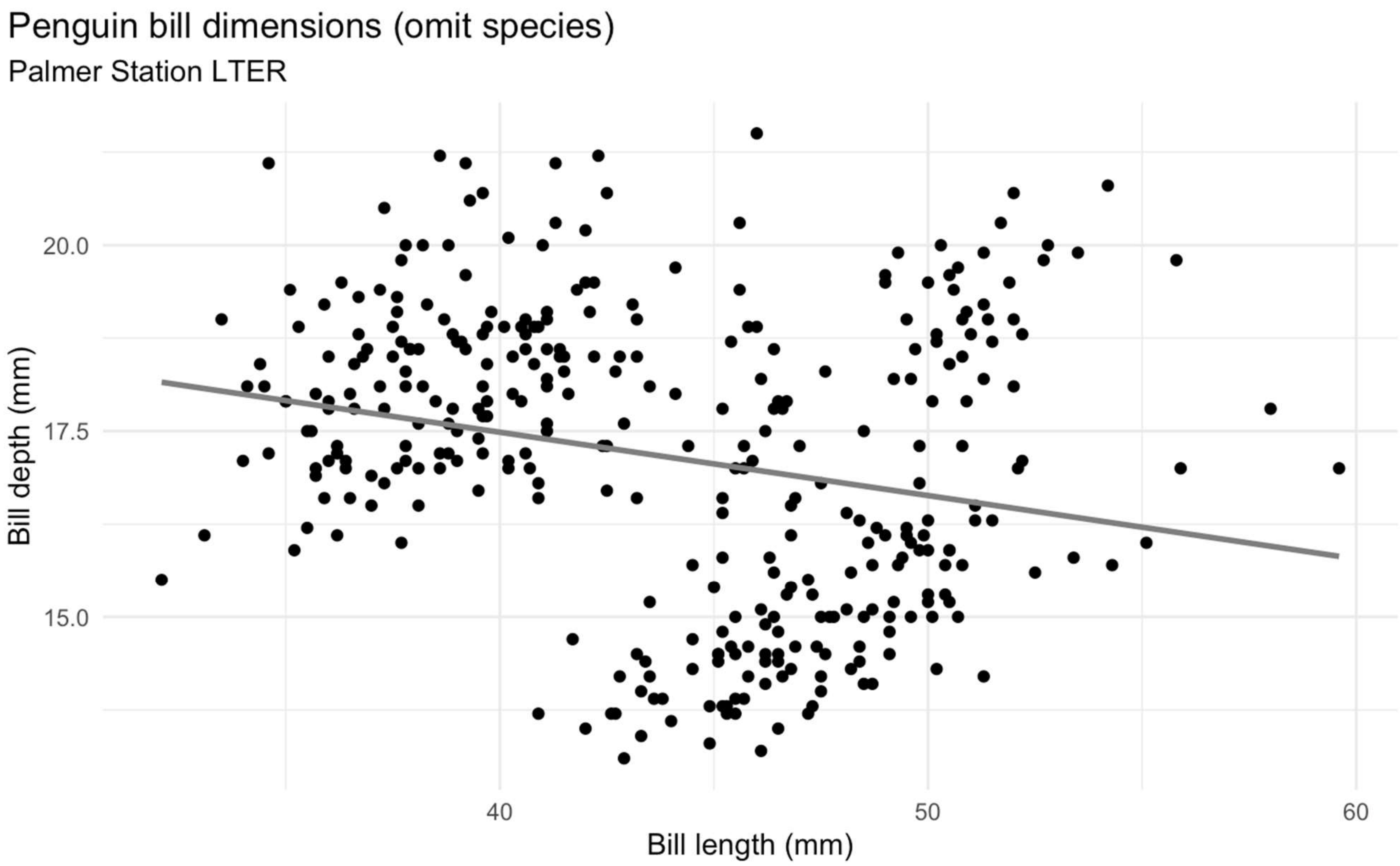
Which of these hypotheses do we test?

Which hypothesis do we test first?

```
penguins %>% select(bill_length_mm,
bill_depth_mm, species)

bill_length_mm bill_depth_mm species
         <dbl>         <dbl> <fct>
 1        39.1          18.7 Adelie
 2        39.5          17.4 Adelie
 3        40.3          18   Adelie
 4        NA            NA   Adelie
 5        36.7          19.3 Adelie
 6        39.3          20.6 Adelie
 7        38.9          17.8 Adelie
 8        39.2          19.6 Adelie
 9        34.1          18.1 Adelie
10        42            20.2 Adelie
```

Penguin bill length decreases with increasing bill depth

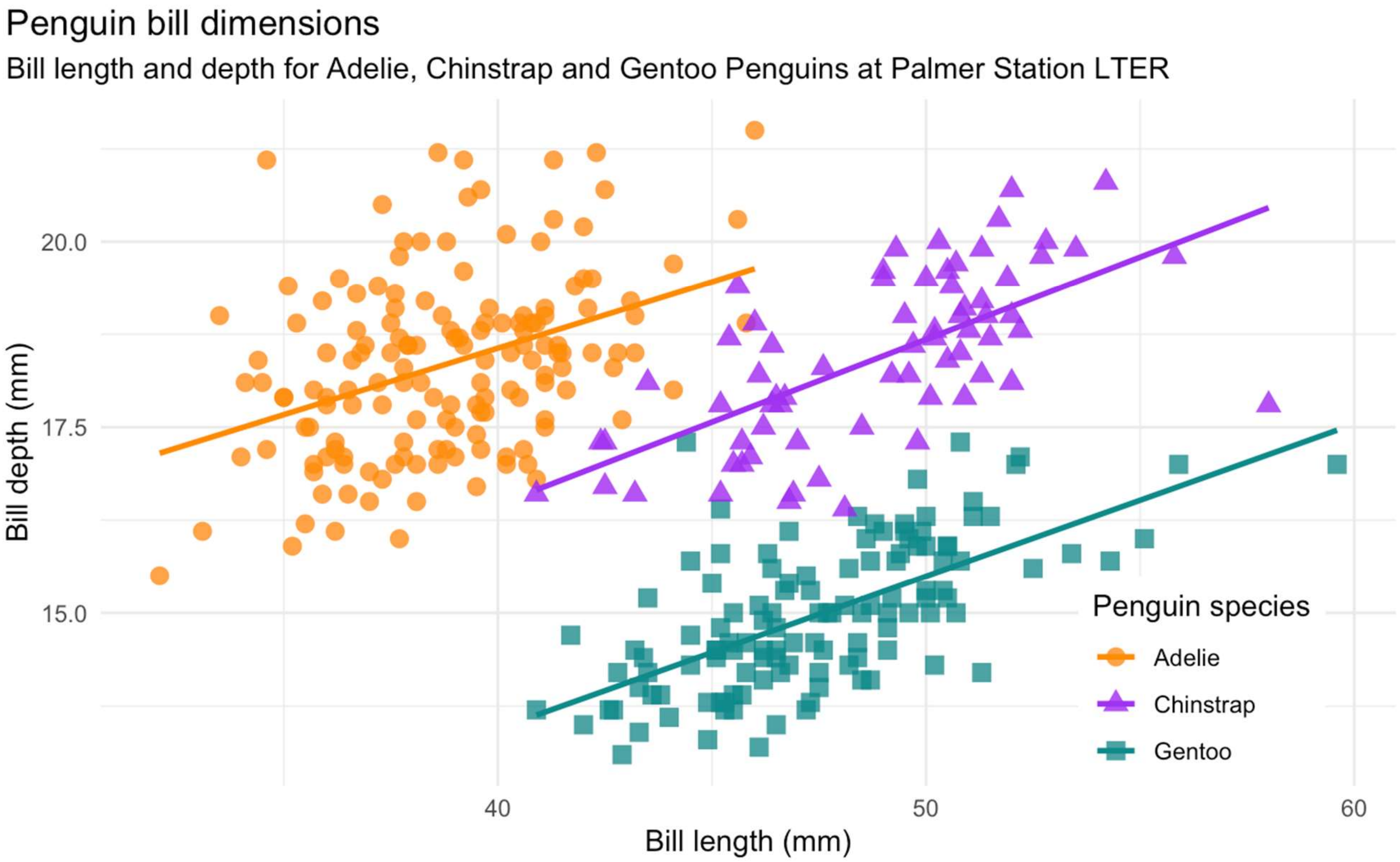


Penguin bill dimensions (omit species)

Palmer Station LTER

Penguin bill length increases with increasing bill depth

Does it look like there is an interaction effect here?



If you omit an important predictor variable, your interpretation can be completely wrong.

Which means a really important part of doing regression is THINKING REALLY HARD about what data you’ll need to collect/record in order to avoid omitted variable bias.

The other side of the coin – Including irrelevant variables/overfitting

- Estimates will be unbiased but perhaps inefficient (greater variances, SEs)

- Hypothesis testing, CI results are valid

- Greater sample size can help reduce inconsistency and selection bias issues

*So: **it’s better to INCLUDE an irrelevant predictor variable that to OMIT a relevant one.***
BUT – not a substitute for having a hypothesis

Overview of Regression Pitfalls/Considerations:

- Omitted Variable Bias (OVB) *(experimental design & conceptual understanding)*
- Inclusion of erroneous variables (not as bad) *(experimental design & conceptual understanding)*
- Assumption violation *(pretty robust, but check diagnostic plots in R for normality & variances)*

Next time

- Multicollinearity

Next time

- Poisson & Binomial models
- Model fitting – balancing between over and underfitting
- Testing & Avoiding Multicollinearity
- Stats test essential checklists