Quantitative methods - solutions to exercises

Katrín Björnsdóttir & Camila Pacheco Riaño

2025-04-15

1. Hierarchical linear model exercise

Libraries

Remember that you have to load the packages each time you start a new R session

```
library(lme4)
```

```
## Loading required package: Matrix
```

```
library(tidyverse)
```

```
## ── Attaching core tidyverse packages ──────────────────────── tidyverse 2.0.0 ──
## ✓ dplyr     1.1.4     ✓ readr     2.1.5
## ✓ forcats   1.0.0     ✓ stringr   1.5.1
## ✓ ggplot2   3.5.1     ✓ tibble    3.2.1
## ✓ lubridate 1.9.4     ✓ tidyr     1.3.1
## ✓ purrr     1.0.4
```

```
## ── Conflicts ────────────────────────────────────────── tidyverse_conflicts() ──
## ✕ tidyr::expand() masks Matrix::expand()
## ✕ dplyr::filter() masks stats::filter()
## ✕ dplyr::lag()    masks stats::lag()
## ✕ tidyr::pack()   masks Matrix::pack()
## ✕ tidyr::unpack() masks Matrix::unpack()
## ℹ Use the conflicted package (<http://conflicted.r-lib.org/>) to force all conflicts to become
errors
```

```
library(ggeffects)
```

Load the data

```
itex <- read_csv("Data/ITEX_diversity_data.csv") %>%
  dplyr::select(-...1) # this code takes out the first column that was called "...1", it's not n
ecessary though, it just makes the dataframe cleaner I think
```

Lets look at the data to make sure it looks okay

```
head(itex)
```

```
## # A tibble: 6 × 10
##   SITE      SUBSITE            PLOT   YEAR TRTMT Latitude WarmQuarterTemp SppRich
##   <chr>     <chr>              <chr> <dbl> <chr>    <dbl>           <dbl>   <dbl>
## 1 ALEXFIORD ALEXFIORD:CASSIO… Cas.…  2007 CTL       78.9            25.6       9
## 2 ALEXFIORD ALEXFIORD:CASSIO… Cas.…  2007 CTL       78.9            25.6       9
## 3 ALEXFIORD ALEXFIORD:CASSIO… Cas.…  2007 CTL       78.9            25.6       7
## 4 ALEXFIORD ALEXFIORD:CASSIO… Cas.…  2007 CTL       78.9            25.6       8
## 5 ALEXFIORD ALEXFIORD:CASSIO… Cas.…  2007 CTL       78.9            25.6       6
## 6 ALEXFIORD ALEXFIORD:CASSIO… Cas.…  2007 CTL       78.9            25.6      10
## # ℹ 2 more variables: `row_number()` <dbl>, PlotTemp <dbl>
```

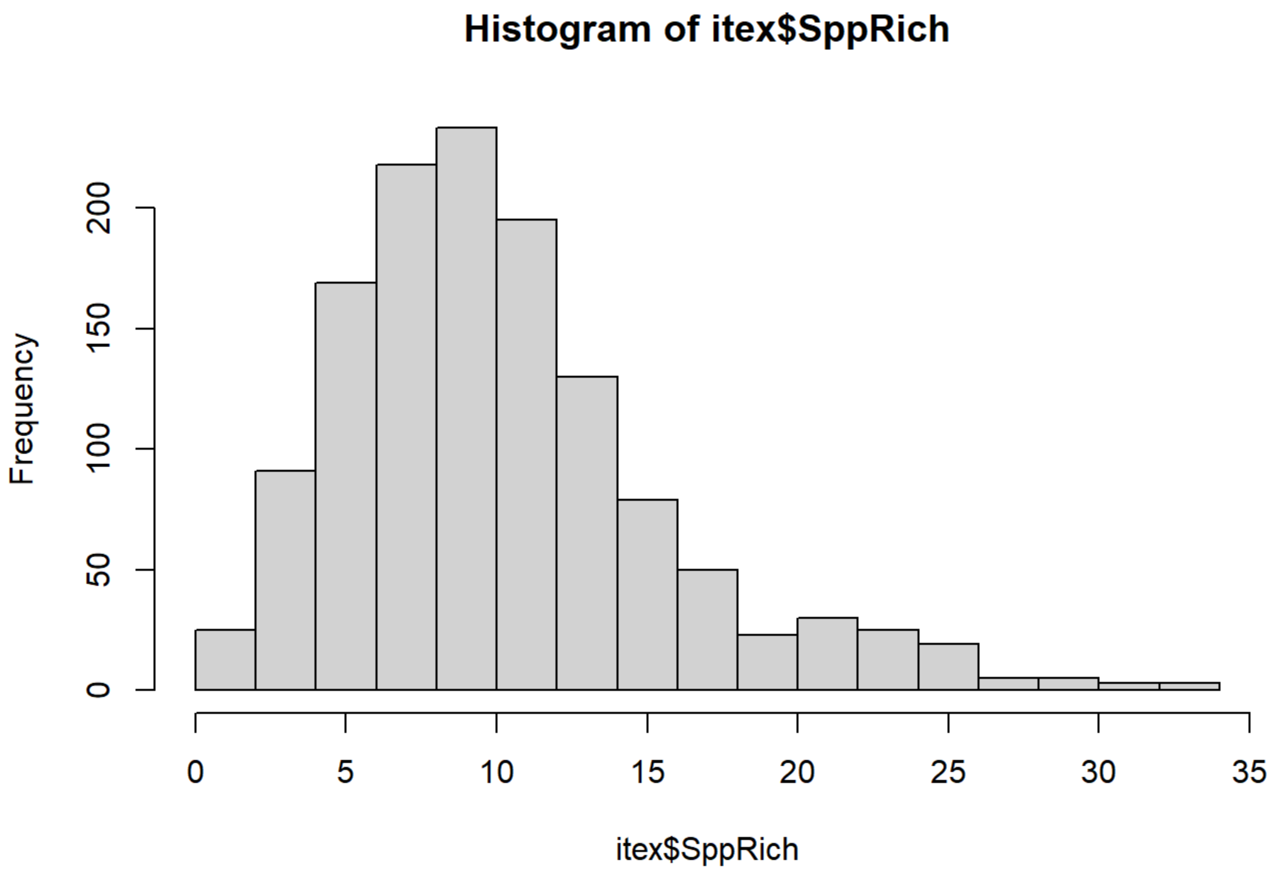
Exercise

What is the relationship between diversity and temperature across sites (i.e. across latitude)?

```
itex <- itex %>%
  mutate(TempC = (WarmQuarterTemp - 30)/2) #change temperature from F to °C (optional)
```

mutate is used to make a new column, in this case we are making a new column called “TempC” by subtracting 30 from the column “PlotTemp” and dividing by 2.

```
hist <- hist(itex$SppRich) # this shows us the distribution of our response variable
```



By exploring the relationship between diversity and temperature across sites we could start with a simple linear model.

```
itex.lm <- lm(SppRich ~ TempC, data = itex)
summary(itex.lm)
```

```
##
## Call:
## lm(formula = SppRich ~ TempC, data = itex)
##
## Residuals:
##      Min       1Q   Median       3Q      Max
## -10.6490  -3.5782  -0.8555   2.3749  22.3021
##
## Coefficients:
##             Estimate Std. Error t value Pr(>|t|)
## (Intercept)  8.95220    0.20016   44.73   <2e-16 ***
## TempC        0.14700    0.01179   12.47   <2e-16 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 5.095 on 1292 degrees of freedom
##   (9 observations deleted due to missingness)
## Multiple R-squared:  0.1074, Adjusted R-squared:  0.1067
## F-statistic: 155.5 on 1 and 1292 DF,  p-value: < 2.2e-16
```

From the summary output you can read plenty, but I'll explain a few for now. -

- The estimate is the slope of the relationship (1.7669)
- Pr is the p value from the model - in our case it is significant since it is lower than 0.05.
- Adjusted R-squared tells us how much variation was explained by the model - in our case 0.1067 or 10%.

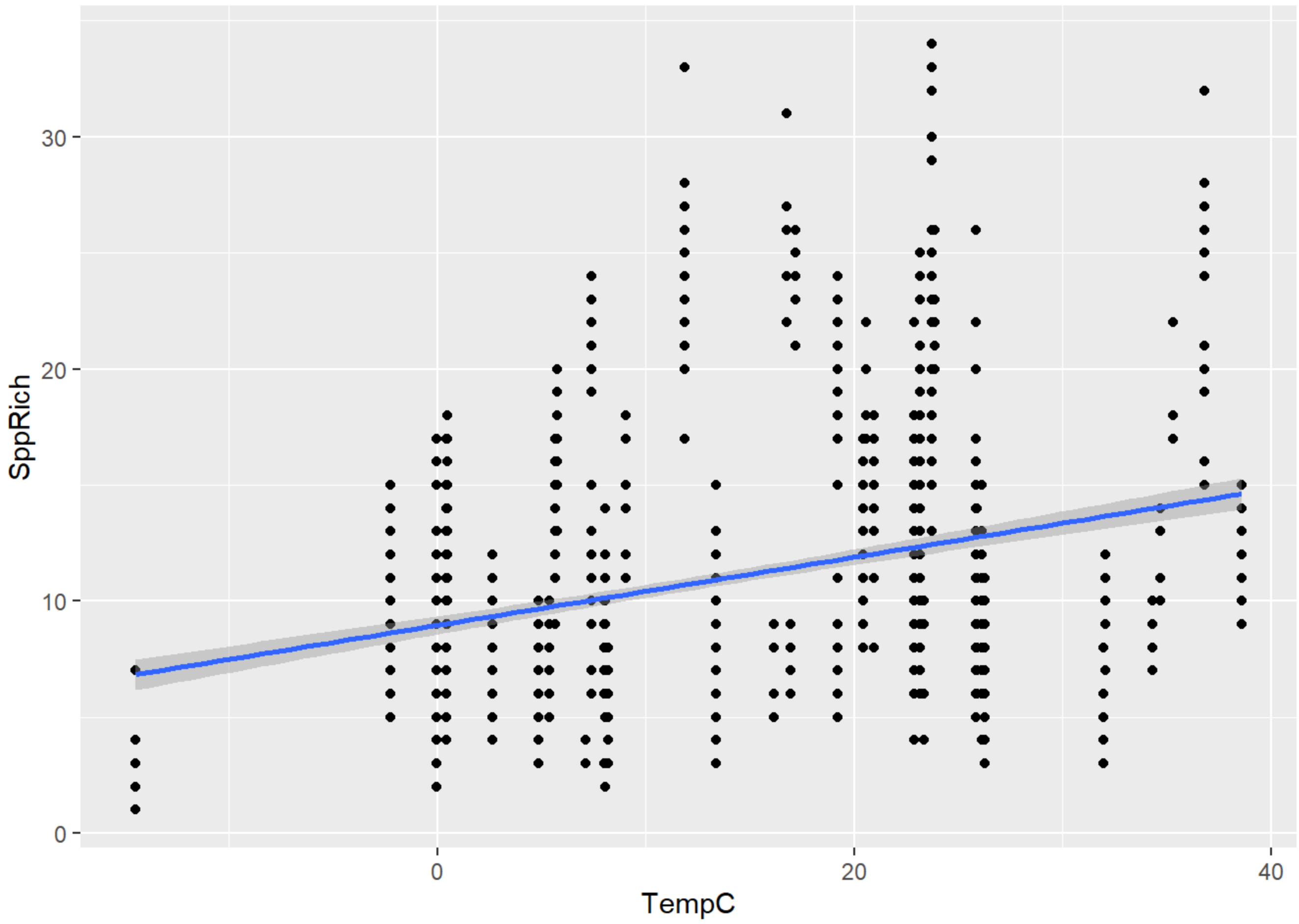
Now lets plot this relationship

```
(prelim_plot <- ggplot(itex, aes(x = TempC, y = SppRich)) +
    geom_point() +
    geom_smooth(method = "lm"))
```

```
## `geom_smooth()` using formula = 'y ~ x'
```

```
## Warning: Removed 9 rows containing non-finite outside the scale range
## (`stat_smooth()`).
```

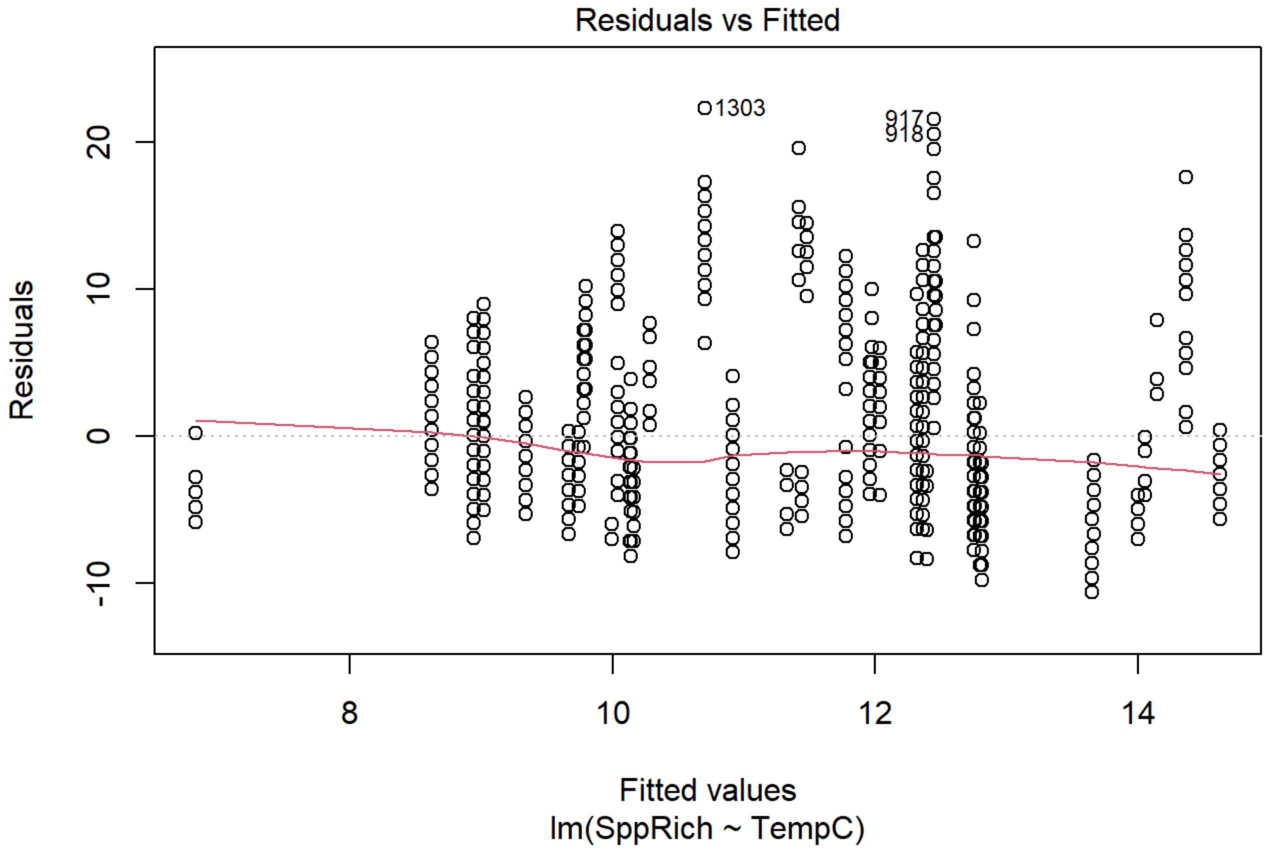
```
## Warning: Removed 9 rows containing missing values or values outside the scale range
## (`geom_point()`).
```



Notice that we get a warming message … wonder what is causing that?

To evaluate how good the model fit was we can look at the residuals (the variation that was not explained by the model) against the fitted values. This is done to detect non-linearity, unequal error variances and outliers. When a linear regression model is suitable for a data set, then the residuals are more or less randomly distributed around the 0 line.

```
plot(itex.lm, which = 1)  # not perfect... but we'll go with it
```

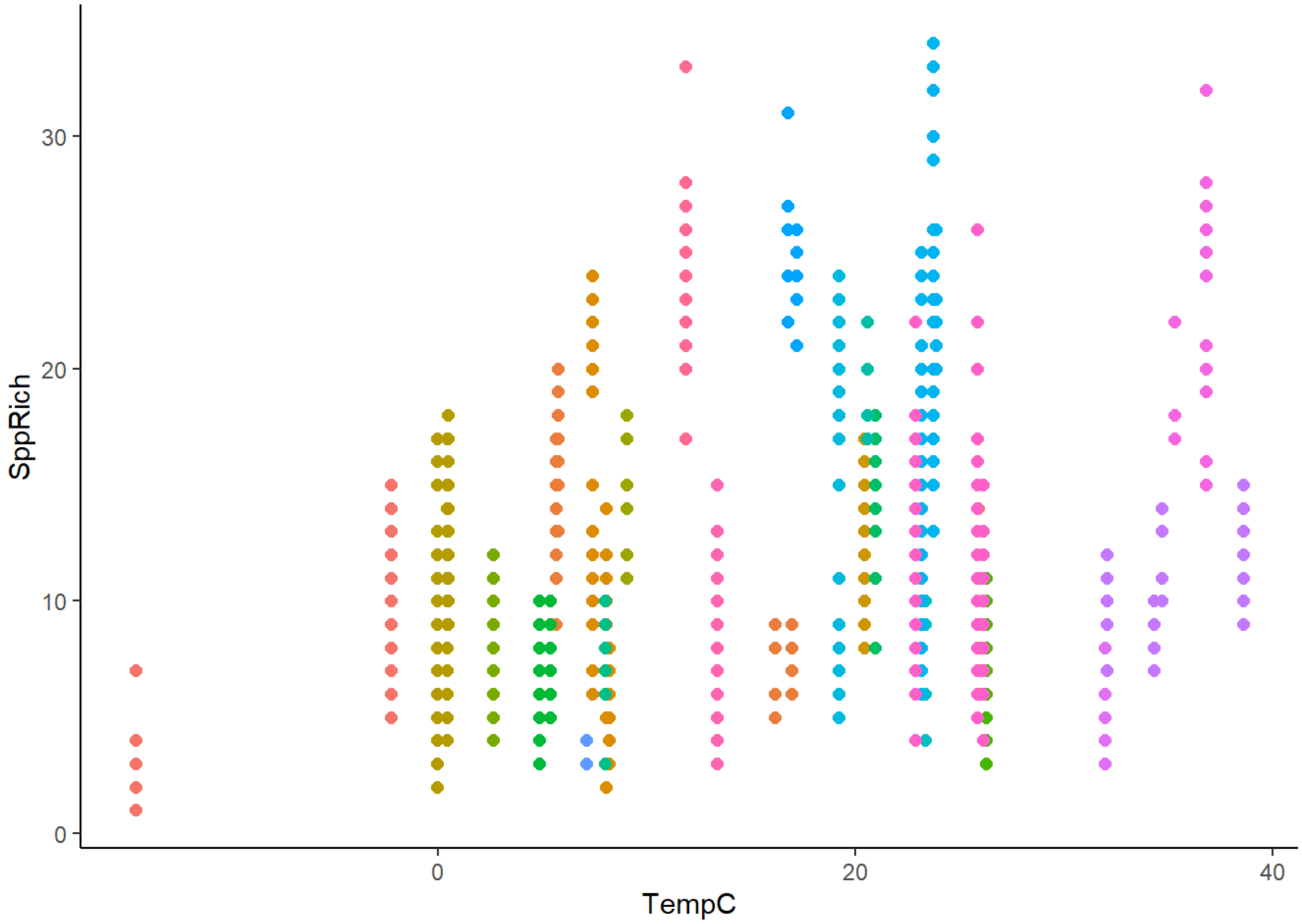


Now lets explore this relationship a bit more, lets add colors based on sites to see if we notice any pattern.

sidenote : in ggplot you can add colors by including “color = SITE” within the aesthetic

```
(colour_plot <- ggplot(itex, aes(x =TempC, y = SppRich, color = SITE)) +
    geom_point(size = 2) +
    theme_classic() +
    theme(legend.position = "none"))
```

```
## Warning: Removed 9 rows containing missing values or values outside the scale range
## (`geom_point()`).
```



Can we see any pattern? … Yes I think so. The colors which represent different sites are in some cases clustered together.

Because of the nested structure of our data, i.e. plots within subsites within sites we might want to think about using a **linear mixed model**. Then we need to think about what we want to include as random effects. Because we expect plots within SITE and SUBSITE to be more similar to each other, we will add SITE and SUBSITE as random effect

```
mixed.lmer <- lmer(SppRich ~ TempC + (1|SITE/SUBSITE), data = itex)
summary(mixed.lmer)
```

```
## Linear mixed model fit by REML ['lmerMod']
## Formula: SppRich ~ TempC + (1 | SITE/SUBSITE)
##    Data: itex
##
## REML criterion at convergence: 6380.5
##
## Scaled residuals:
##     Min      1Q  Median      3Q     Max
## -3.3464 -0.5508 -0.0273  0.4826  5.6335
##
## Random effects:
##  Groups       Name        Variance Std.Dev.
##  SUBSITE:SITE (Intercept) 21.046   4.588
##  SITE         (Intercept) 17.059   4.130
##  Residual                  6.334   2.517
## Number of obs: 1294, groups:  SUBSITE:SITE, 82; SITE, 23
##
## Fixed effects:
##             Estimate Std. Error t value
## (Intercept)  9.66769    1.71768   5.628
## TempC        0.14259    0.08659   1.647
##
## Correlation of Fixed Effects:
##       (Intr)
## TempC -0.774
```

Lets first look at if the summary output corresponds to the structure of the data.

From the output we can see that the model was fitted with: 1294 obs., 82 SUBSITES and 23 SITES

see this line in the summary output - Number of obs: 1294, groups: SUBSITE:SITE, 82; SITE, 23

Hmmmm… that’s not quite what we have is it?

```
nrow(itex) # we have 1303 observation
```

```
## [1] 1303
```

```
length(unique(itex$SITE)) # 24 sites
```

```
## [1] 24
```

```
length(unique(itex$SUBSITE)) # 83 subsites
```

```
## [1] 83
```

Something is not quite right here. Do you remember how we got a warning message earlier in the script when we were plotting the relationship?

The problem we have is that one of the sites has missing values that we should remove before the analysis

Lets look for missing values then

```
sum(is.na(itex$SppRich)) # no NAs in SppRich, how about TempC_scaled?
```

```
## [1] 0
```

```
sum(is.na(itex$TempC)) # yes, here we have 9 missing values
```

```
## [1] 9
```

```
itex <- itex %>%
  filter(!is.na(TempC)) # here we filter out all NA values within TempC_scaled
```

sidenote: adding a “!” infront will remove values when using the filter function

Alright lets try it again.

```
nrow(itex) # now we have 1294 observation
```

```
## [1] 1294
```

```
length(unique(itex$SITE)) # 23 sites
```

```
## [1] 23
```

```
length(unique(itex$SUBSITE)) # 82 subsites
```

```
## [1] 82
```

Great now we can continue.

Run the model again with the fixed dataset.

```
mixed.lmer <- lmer(SppRich ~ TempC + (1|SITE/SUBSITE), data = itex)
summary(mixed.lmer)
```

```
## Linear mixed model fit by REML ['lmerMod']
## Formula: SppRich ~ TempC + (1 | SITE/SUBSITE)
##    Data: itex
## 
## REML criterion at convergence: 6380.5
## 
## Scaled residuals: 
##     Min      1Q  Median      3Q     Max 
## -3.3464 -0.5508 -0.0273  0.4826  5.6335 
## 
## Random effects:
##  Groups       Name        Variance Std.Dev.
##  SUBSITE:SITE (Intercept) 21.046   4.588   
##  SITE         (Intercept) 17.059   4.130   
##  Residual                  6.334   2.517   
## Number of obs: 1294, groups:  SUBSITE:SITE, 82; SITE, 23
## 
## Fixed effects:
##             Estimate Std. Error t value
## (Intercept)  9.66769    1.71768   5.628
## TempC        0.14259    0.08659   1.647
## 
## Correlation of Fixed Effects:
##       (Intr)
## TempC -0.774
```

The summary output for linear mixed models in the lme4 package don’t show p-values.This is a conscious choice made by the authors of the package, as there are many problems with p-values. Please be careful when it comes to evaluating the significance of your model. Focus on your question and avoid plugging in and dropping variables from a model haphazardly until you make something “significant”. Always choose variables based on biology/ecology and avoid adding in all possible variables that you have access to (i.e. don’t overfit). Remember that as a rule of thumb, you need 10 times more data than parameters you are trying to estimate. There are some common different methods used for model selection which we won’t go into here but take a look at this thread if you are interested: https://stats.stackexchange.com/questions/95054/how-to-get-an-overall-p-value-and-effect-size-for-a-categorical-factor-in-a-mi (https://stats.stackexchange.com/questions/95054/how-to-get-an-overall-p-value-and-effect-size-for-a-categorical-factor-in-a-mi)

For simplification, we’ll use the stargazer package that gives us a summary output table from our model.

```
#remember to install it first if you haven't already - then load!
library(stargazer)
```

```
stargazer(mixed.lmer,
          digits = 3,
          type="text",
          star.cutoffs = c(0.05, 0.01, 0.001),
          digit.separator = "")
```

```
##
## =================================================
##                           Dependent variable:
##                      -----------------------------
##                                 SppRich
## -------------------------------------------------
## TempC                           0.143
##                                (0.087)
##
## Constant                      9.668***
##                                (1.718)
##
## -------------------------------------------------
## Observations                    1294
## Log Likelihood                -3190.238
## Akaike Inf. Crit.             6390.476
## Bayesian Inf. Crit.           6416.303
## =================================================
## Note:               *p<0.05; **p<0.01; ***p<0.001
```

From this table we see that there is not a significant relationship with temperature and species richness across sites. If there would have been a significant relationship, there would have been an asterick/s after the 1.714 value (the slope of relationship).

Now for fun, lets visualize the new model to compare with our original linear model

Extract the prediction dataframe.

```
pred.mm <- ggpredict(mixed.lmer, terms = c("TempC"))  # this gives an overall predictions for th
e model.
```

Now lets try to plot the predictions.

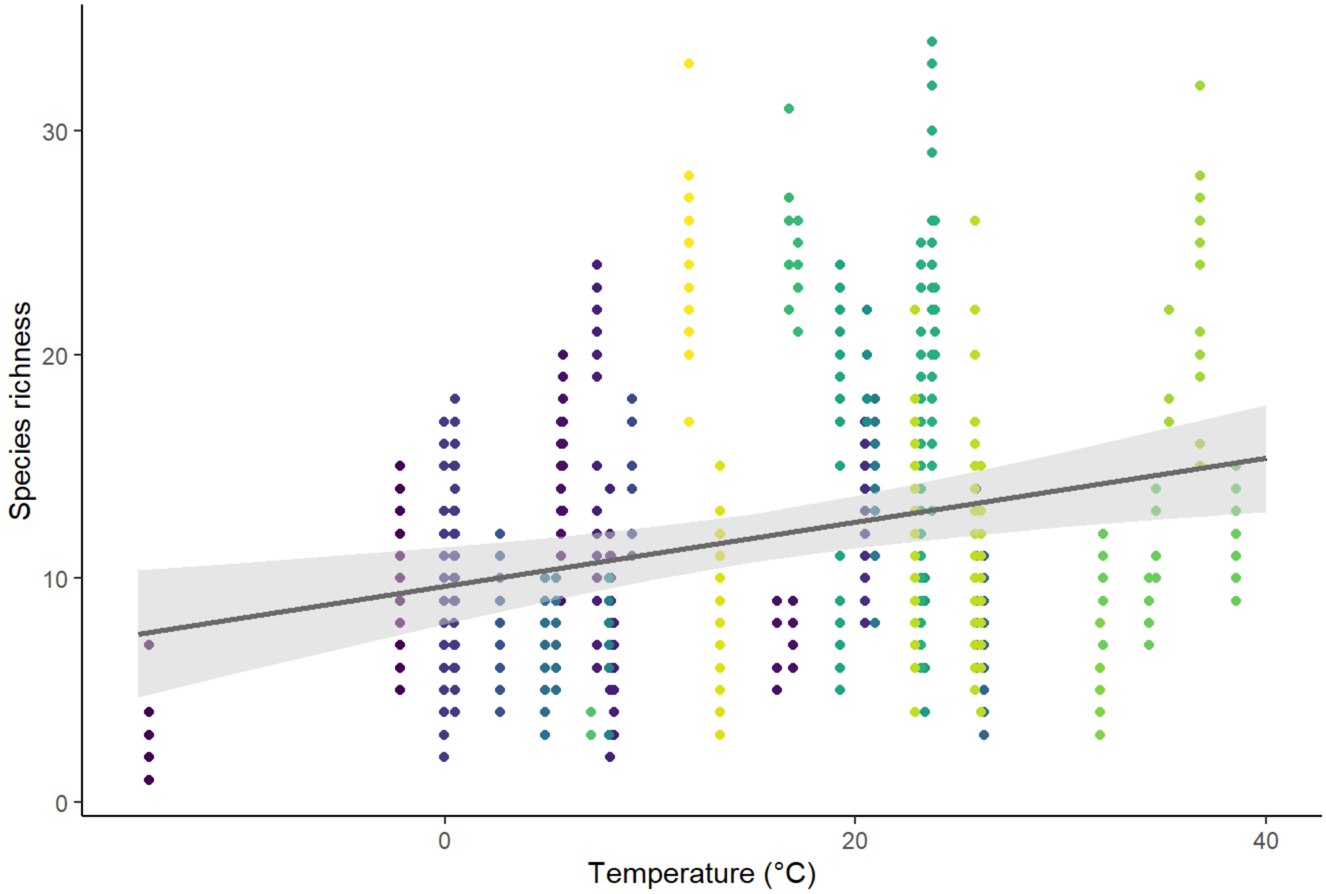
```
#install.packages("viridis") #install this package to be able to add nice colors
```

Sidenote: Only install this if you haven't done that previously, otherwise go ahead and load it.

```
library(viridis)
```

```
## Loading required package: viridisLite
```

```
(ggplot(pred.mm) +
    geom_point(data = itex,
               aes(x = TempC, y = SppRich, colour = SITE)) + # adding the raw data (scaled values) as points
    geom_line(aes(x = x, y = predicted), linewidth = 1, color = "black") +  # this is the slope
    geom_ribbon(aes(x = x, ymin = predicted - std.error, ymax = predicted + std.error),
                fill = "lightgrey", alpha = 0.5) +  # this is the error band
    labs(x = "Temperature (°C)", y = "Species richness",
         title = "") +
    scale_color_viridis(discrete = TRUE, option = "viridis") + #this will add nice colors to the plot from the viridis package
    theme_classic() +
    theme(legend.position = "none") # here we are removing the legend
)
```



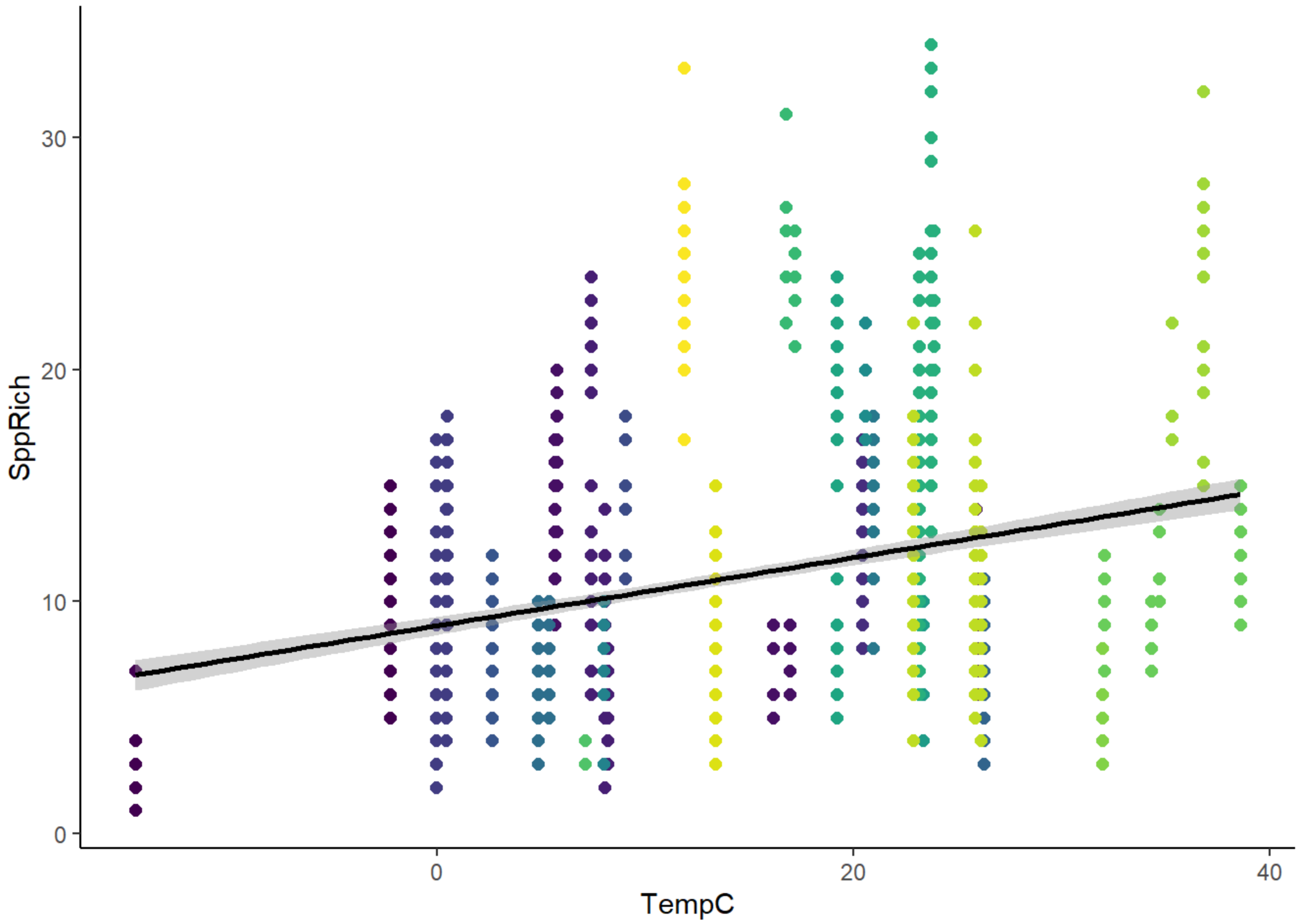
Now we can compare the two models to see how they differ visually.

```
summary(itex.lm) #this is the output from the linear model
```

```
##
## Call:
## lm(formula = SppRich ~ TempC, data = itex)
##
## Residuals:
##      Min       1Q   Median       3Q      Max
## -10.6490  -3.5782  -0.8555   2.3749  22.3021
##
## Coefficients:
##             Estimate Std. Error t value Pr(>|t|)
## (Intercept)  8.95220    0.20016   44.73   <2e-16 ***
## TempC        0.14700    0.01179   12.47   <2e-16 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 5.095 on 1292 degrees of freedom
##   (9 observations deleted due to missingness)
## Multiple R-squared:  0.1074, Adjusted R-squared:  0.1067
## F-statistic: 155.5 on 1 and 1292 DF,  p-value: < 2.2e-16
```

```
(colour_plot <- ggplot(itex) +
    geom_point(size = 2, aes(x = TempC, y = SppRich, color = SITE)) +
    geom_smooth(method = "lm", aes(x = TempC, y = SppRich), color = "black") +
    scale_color_viridis(discrete = TRUE, option = "viridis") +
    theme_classic() +
    theme(legend.position = "none"))
```

```
## `geom_smooth()` using formula = 'y ~ x'
```



Can you make sense of the difference? Notice how the slopes are very similar (1.7669 for the linear and 1.714 for the linear mixed model). What is different is the variation across sites. By adding a random effect to our model, we no longer detected a positive relationship with species richness and temperature across sites because of the increased uncertainty.

Now this is it for linear mixed modelling - hope you've learned something :-)

Remember to check out the Coding Club's INTRODUCTION TO LINEAR MIXED MODELS, find it here: https://ourcodingclub.github.io/tutorials/mixed-models/ (https://ourcodingclub.github.io/tutorials/mixed-models/)

and this is also a great source to understand how to work with colors in ggplot: https://r-graph-gallery.com/ggplot2-color.html (https://r-graph-gallery.com/ggplot2-color.html)

Don't hesitate to contact us if you are having problems with running the script.

katrin.bjornsdottir@bioenv.gu.se (mailto:katrin.bjornsdottir@bioenv.gu.se)

camila.pacheco.riano@bioenv.gu.se (mailto:camila.pacheco.riano@bioenv.gu.se)

Good luck :-)