ENTMLGY 6702 Entomological Techniques and Data Analysis

R Activity 7: Multiple Linear Regression (MLR)

Due: 10/24/2023

We have been using simple linear regression (SLR) to evaluate bivariate relationships, such as quantifying the effect of one predictor on our response variable (i.e., the effect of X on Y, Y as a function of X, or, in R speak, `lm(Y~X)`):

$Y_i = \beta_0 + \beta_1 X_1 + \epsilon_i$

$Plant\ height \sim\ DBH$

In multiple linear regression (MLR), we are investigating the explanatory power of two or more predictors, also referred to as "covariates" (note the subscripts for X are different, meaning these variables represent different predictors):

$Y_i = \beta_0 + \beta_1 X_1 + \beta_2 X_2 + \epsilon_i$

$Plant\ height \sim\ DBH + soil\ nitrogen$

1 Assumptions and collinearity

The same assumptions for SLR ($Y \sim X_1$) apply to MLR ($Y \sim X_1 + X_2$) but there are some additional assumptions in MLR. If you fit a MLR model, you should also check for **collinearity** (aka multicollinearity): correlation between your predictors. If two predictors are highly correlated, they end up explaining the same "type" of variation in your response. For example, if you were trying to predict weight of puppies using age and height, the strong relationship between puppy age and height would cause a problem; when collinearity is present, strange things can happen in the model. For example, the estimate of a slope for a given predictor might flip from positive to negative when that predictor is fit alone (= SLR) vs. when it is fit with a second predictor (= MLR) with which it is strongly correlated.

In MLR, we interpret coefficients as follows: "holding all else equal, a one unit increase in X_1 was associated with a B_1 unit increase in Y." That is part of the reason collinearity causes problems. If predictors X_1 and X_2 are highly correlated, it is difficult to "hold X_2 equal" while estimating what happens to the expected value of Y with a one unit change in X_1.

2 Fitting a MLR model

I will use the `CO2` data set from the `datasets` package in R. This data set reports changes in CO_2 uptake (μmol/m^2 *sec*) with (i) `conc` (ambient CO_2 in mL/L), (ii) `Type` (origin of the plant: `Quebec` or `Mississippi`), and (iii) `Treatment` (`chilled` or `nonchilled` overnight before the experiment).

Please note a few things here. These models (`Y ~ continuous X + categorical Z`) are often referred to as analysis of covariance, or ANCOVA, whether an interaction term is included or not. We are going to ignore the repeated measures and `Treatment` and just look at the effects of `conc` and `Type`. There is no interaction term in the below model, and so we say we are just including "main effects" (a bit more on this below). I also provide a graphical depiction of the model, in which you should see that the fit lines corresponding to each level of `Type` are perfectly parallel (just FYI, MLR models with >3 predictors are often presented in tables rather than graphs).

```
library(datasets)
fit_plants_1_nointeraction <- lm(uptake~conc+Type,data=CO2)
anova(fit_plants_1_nointeraction)
```

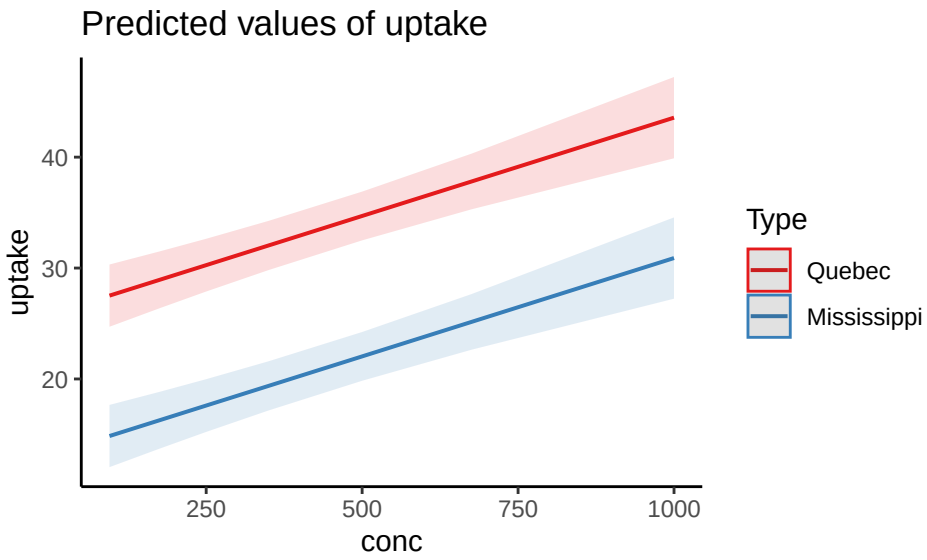
```
## Analysis of Variance Table
##
## Response: uptake
##           Df Sum Sq Mean Sq F value    Pr(>F)
## conc       1 2285.0  2285.0  45.627 1.997e-09 ***
## Type       1 3365.5  3365.5  67.204 3.061e-12 ***
## Residuals 81 4056.4    50.1
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

```
summary(fit_plants_1_nointeraction)
```

```
##
## Call:
## lm(formula = uptake ~ conc + Type, data = CO2)
##
## Residuals:
##      Min       1Q   Median       3Q      Max
## -18.2145  -4.2549   0.5479   5.3048  12.9968
##
## Coefficients:
##                   Estimate Std. Error t value Pr(>|t|)
## (Intercept)      25.830052   1.579918  16.349  < 2e-16 ***
## conc              0.017731   0.002625   6.755 2.00e-09 ***
## TypeMississippi -12.659524   1.544261  -8.198 3.06e-12 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 7.077 on 81 degrees of freedom
## Multiple R-squared:  0.5821, Adjusted R-squared:  0.5718
## F-statistic: 56.42 on 2 and 81 DF,  p-value: 4.498e-16
```

In our model with no interaction, our interpretation would be that plants from `Quebec` had higher uptake than plants from `Mississippi` AND that the differences in `uptake` were constant across all values of `conc` (the lines are parallel!). Said another way, the relationship (i.e., slope) between `uptake` and `conc` does not change with plant `Type` - the intercept does change with plant `Type`, however (that is apparent from looking at the graph).

```
library(tidyverse)
library(sjPlot)
library(sjmisc)
plot_model(fit_plants_1_nointeraction, type = "pred", terms = c("conc", "Type"))+
  theme_classic()
```



3 Interactions

Interactions between our predictors indicate whether the effect of one predictor (X_1) on our response variable (Y) is influenced by another predictor (X_2). Note this is different from collinearity, because when fitting an interaction our goal is actually to quantify how one predictor affects the relationship between our response and other predictor. In R, we code interactions using an asterisk, "`*`". In the case of `uptake`, we might be interested to know if the relationship between `uptake` and `conc` changes between plants from `Quebec` vs. `Mississippi` (i.e., levels of `Type`).

Note that we always include the so-called main effects when fitting interaction terms: you would not fit a variable in an interaction term without also fitting that variable alone (i.e., as a main effect). Indeed, you will notice below that `conc*Type` forces R to provide estimates of main effects for `conc` and `Type` in the `summary()`.

```
fit_plants_1_interaction <- lm(uptake~conc*Type,data=CO2)
anova(fit_plants_1_interaction)
```

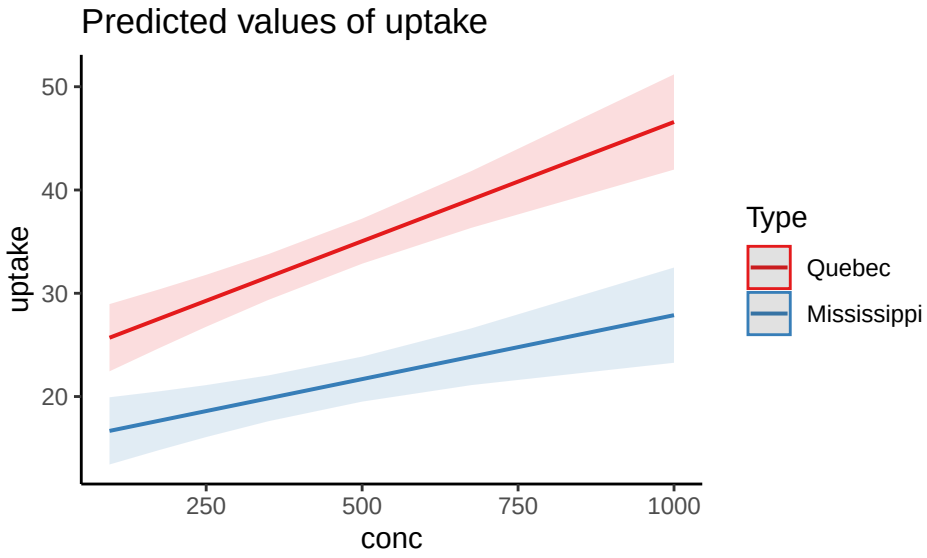
```
## Analysis of Variance Table
##
## Response: uptake
##           Df Sum Sq Mean Sq F value    Pr(>F)
## conc       1 2285.0  2285.0 47.4995 1.143e-09 ***
## Type       1 3365.5  3365.5 69.9614 1.560e-12 ***
## conc:Type  1  208.0   208.0  4.3238   0.04079 *
## Residuals 80 3848.4    48.1
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

```
summary(fit_plants_1_interaction)
```

```
##
## Call:
## lm(formula = uptake ~ conc * Type, data = CO2)
##
## Residuals:
##      Min       1Q   Median       3Q      Max
## -16.3956  -5.5250  -0.1604   5.5724  12.0072
##
## Coefficients:
##                       Estimate Std. Error t value Pr(>|t|)
## (Intercept)          23.503038   1.910531  12.302  < 2e-16 ***
## conc                  0.023080   0.003638   6.344 1.25e-08 ***
## TypeMississippi      -8.005495   2.701899  -2.963  0.00401 **
## conc:TypeMississippi -0.010699   0.005145  -2.079  0.04079 *
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 6.936 on 80 degrees of freedom
## Multiple R-squared:  0.6035, Adjusted R-squared:  0.5887
## F-statistic: 40.59 on 3 and 80 DF,  p-value: 4.78e-16
```

In the graph below, note the lines are no longer parallel! With the interaction term, we are now evaluating whether the effect of `conc` on `uptake` changes with plant `Type`. In this case, we would conclude that `Type` influenced the `uptake ~ conc` relationship because of the `conc:Type` line in the above anova table ($F_{1,80} = 4.32, p = 0.0408$). In the next section, I go into more detail on how to interpret `summary()` output of models with interactions.

```
plot_model(fit_plants_1_interaction, type = "pred", terms = c("conc", "Type"))+
  theme_classic()
```



4 Multiple ANOVA (MANOVA)

We have worked with one-way ANOVAs to quantify variation in a continuous response variable as a function of a single predictor that had multiple levels (e.g., plant growth as a function of "fertilizer", where fertilizer had three different levels or fertilizer types). In the preceding section, we used ANOVA to evaluate the effects of a categorical and a continuous predictor on a continuous response. ANOVA is quite flexible/powerful, and we can also evaluate the effects of multiple categorical predictors - that each have two or more levels - on a continuous response.

It is very typical to fit an interaction term in such models, but that will depend on the specific goals of your analysis. Below we are fitting a two-way interaction (e.g., `A*B`), but you can fit multi-way interactions (`A*B*C`) if you have a compelling biological reason, but be warned that the more interactions the more difficult the model is to interpret. Also note that instead of `Treatment+Type+Treatment*Type`, I could have written `Treatment*Type` and the same model would have been fit (please re-read the previous section if that coding option seems strange).

```
fit_manova <- lm(uptake~Treatment+Type+Treatment*Type,data=CO2)
anova(fit_manova)
```

```
## Analysis of Variance Table
##
## Response: uptake
##                Df Sum Sq Mean Sq F value    Pr(>F)
## Treatment       1  988.1   988.1 15.4164 0.0001817 ***
## Type            1 3365.5  3365.5 52.5086 2.378e-10 ***
## Treatment:Type  1  225.7   225.7  3.5218 0.0642128 .
## Residuals      80 5127.6    64.1
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

```
summary(fit_manova)
```

```
##
## Call:
## lm(formula = uptake ~ Treatment + Type + Treatment * Type, data = CO2)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -22.452  -3.624   2.167   5.773  10.648
##
## Coefficients:
##                                 Estimate Std. Error t value Pr(>|t|)
## (Intercept)                       35.333      1.747  20.225  < 2e-16 ***
## Treatmentchilled                  -3.581      2.471  -1.449 0.151141
## TypeMississippi                   -9.381      2.471  -3.797 0.000284 ***
## Treatmentchilled:TypeMississippi  -6.557      3.494  -1.877 0.064213 .
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 8.006 on 80 degrees of freedom
## Multiple R-squared:  0.4718, Adjusted R-squared:  0.452
## F-statistic: 23.82 on 3 and 80 DF,  p-value: 4.106e-11
```

The interpretation is the same for our other ANOVA models, but note that the `Estimate` column in the above `summary()` output is providing the difference in means of groups compared to a reference level where that reference level (`Intercept`) represents the mean of `uptake` for plants that were `nonchilled` and from `Quebec`. The second line compares the `uptake` of plants that were `chilled` and from `Quebec` to the reference group. The third line is comparing `uptake` of plants that were `nonchilled` and from `Mississippi` to the reference group. The last line, the interaction term (`Treatmentchilled:TypeMississippi`), is comparing `uptake` of plants that were `chilled` and from `Mississippi` to plants that were `nonchilled` and from `Quebec` (which again, is the reference level).

So, to estimate the `uptake` of plants that were `nonchilled` and from `Mississippi`, we would use the following. Note all the numbers are taken from the above `summary()` output and the 0 or 1 indicates "membership" to a treatment (think back to when we covered "dummy" variables):

$$25.952 = 35.333 - (0 \times 3.581) - (1 \times 9.381) - (0 \times 6.557)$$

And to estimate the `uptake` of plants that were `chilled` and from `Mississippi`, we would use the following:
$15.814 = 35.333 - (1 \times 3.581) - (1 \times 9.381) - (1 \times 6.557)$

In a way, our model is just estimating the means of our different treatment groups (nonchilled/Quebec, nonchilled/Mississippi, chilled/Quebec, chilled/Mississippi), which we could also calculate using our raw data (see below). Our formal analyses, however, tell us if any differences in the means are statistically meaningful. Note that to compare all the groups to each other, we would need something like `emmeans` to account for making multiple comparisons (see next section).

```
CO2 %>% group_by(Treatment, Type) %>% summarise(means = round(mean(uptake),3)) %>%
  as.data.frame()
```

```
## `summarise()` has grouped output by 'Treatment'. You can override using the
## `.groups` argument.
```

```
##    Treatment        Type  means
## 1 nonchilled      Quebec 35.333
## 2 nonchilled Mississippi 25.952
## 3    chilled      Quebec 31.752
## 4    chilled Mississippi 15.814
```

5 Multiple comparisons in MLR

We can still conduct multiple comparisons between levels of a factor in MLR/MANOVA frameworks, including when there are interaction terms. However, before doing so, we often need to adjust for the fact that there are other treatments/variables in our model potentially influencing the response variable. . . thankfully, R does that for us! Below are two examples, and the interpretations are the same as other multiple comparisons we have covered. Note that if you have interaction terms, you should probably limit your multiple comparisons to the interaction term alone. That is, interpreting main effects is not easy when you fit interactions.

```
library(emmeans)
```

5.1 No interaction, one categorical and one continous predictor

```
no_interaction_emm <- emmeans(fit_plants_1_nointeraction, ~Type)
pairs(no_interaction_emm)
```

```
##  contrast             estimate   SE df t.ratio p.value
##  Quebec - Mississippi     12.7 1.54 81   8.198  <.0001
```

5.2 Interaction between a categorical and a continous predictor

Note that this is comparing the *slopes* of the `uptake ~ conc` relationship between the `Quebec` group and `Mississippi` group. So, the slope of `uptake ~ conc` is steeper by 0.0107 for the `Quebec` plants.

```
fp_inter_emm <- emtrends(fit_plants_1_interaction, "Type", var = "conc")
pairs(fp_inter_emm)
```

```
##  contrast             estimate      SE df t.ratio p.value
##  Quebec - Mississippi   0.0107 0.00515 80   2.079  0.0408
```

5.3 No interaction, two categorical predictors

Note that we are just looking at `Treatment` here but the output acknowledges the presence of `Type` in the model (`Results are averaged over the levels of: Type`).

```
fit_manova_no_interaction <- lm(uptake~Treatment+Type,data=CO2)
fit_manova_no_interaction_emm <- emmeans(fit_manova_no_interaction, ~Treatment)
pairs(fit_manova_no_interaction_emm)
```

```
##  contrast             estimate   SE df t.ratio p.value
##  nonchilled - chilled     6.86 1.77 81   3.867  0.0002
##
## Results are averaged over the levels of: Type
```

5.4 Interaction between two categorical predictors

```
manova_emm <- emmeans(fit_manova, ~Treatment*Type)
pairs(manova_emm)
```

```
##  contrast                                    estimate   SE df t.ratio p.value
##  nonchilled Quebec - chilled Quebec              3.58 2.47 80   1.449  0.4728
##  nonchilled Quebec - nonchilled Mississippi      9.38 2.47 80   3.797  0.0016
##  nonchilled Quebec - chilled Mississippi        19.52 2.47 80   7.900  <.0001
##  chilled Quebec - nonchilled Mississippi         5.80 2.47 80   2.348  0.0960
##  chilled Quebec - chilled Mississippi           15.94 2.47 80   6.451  <.0001
##  nonchilled Mississippi - chilled Mississippi   10.14 2.47 80   4.103  0.0006
```

```
##
## P value adjustment: tukey method for comparing a family of 4 estimates
```

6 Sequential vs. marginal fitting

The standard R `anova()` command uses something called sequential fitting, or Type I sums of squares (the names Type X sums of squares originated with SAS, another stats software). A full description of the different types of sums squares is beyond the scope of this class, but what this means is that the order in which you list the predictors inside the `lm()` command can influence the statistical output and thus your conclusions. . . wait, what?! To illustrate this point, I'll use a data set on tick abundance.

6.1 Sequential fitting

Please do not worry about biology here, but notice in the output that the sums of squares are changing *just* because we swapped the order of our predictors. Everything else is *exactly* the same.

When `f_YEAR` is listed first:

```
grouseticks$f_YEAR <- as.factor(grouseticks$YEAR)
fit_ex1 <- lm(TICKS ~ f_YEAR + HEIGHT, data=grouseticks)
anova(fit_ex1)
```

```
## Analysis of Variance Table
##
## Response: TICKS
##            Df Sum Sq Mean Sq F value    Pr(>F)
## f_YEAR      2   7050  3524.9  24.995 5.928e-11 ***
## HEIGHT      1   6092  6092.0  43.199 1.550e-10 ***
## Residuals 399  56268   141.0
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

When `HEIGHT` is listed first:

```
fit_ex2 <- lm(TICKS ~ HEIGHT + f_YEAR, data=grouseticks)
anova(fit_ex2)
```

```
## Analysis of Variance Table
##
## Response: TICKS
##            Df Sum Sq Mean Sq F value    Pr(>F)
## HEIGHT      1   7692  7692.2  54.546 8.948e-13 ***
## f_YEAR      2   5450  2724.8  19.321 9.788e-09 ***
## Residuals 399  56268   141.0
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

What the output is doing for `anova(fit_ex1)` is presenting the effect of `f_YEAR` on `TICKS` and THEN presenting the effect of `HEIGHT` on `TICKS` AFTER accounting for the effect of `f_YEAR`. Think of it this way: `f_YEAR` explains some variation in `TICKS` and then we see how good of a job `HEIGHT` does of explaining the remaining variation. The `Residuals` indicate the remaining/leftover variation after accounting for both predictors (= unexplained variation).

The exact reverse is true for the output of `anova(fit_ex2)`, which is presenting the effect of `HEIGHT` on `TICKS` and THEN presenting the effect of `f_YEAR` on `TICKS` AFTER accounting for the effects of `HEIGHT`. The `Residuals` value does not change between analyses, which makes sense: we are using the same two

predictors in both models and they will explain the exact same amount of variability in our response variable (we are just swapping the order in which we fit them).

6.2 Marginal fitting

In practice, we usually - really, always - want to look at the explanatory power of our predictors after accounting for the other predictors in a model (if you have a counter example, I am genuinely curious to hear it!). This process is called marginal fitting, or Type II OR Type III sums of squares.

Again, the nuances of sums of squares are beyond the scope of this class, but here is a short summary. First, remember that interpreting main effects can be very difficult when you have an interaction term in the model (e.g., interpreting effects of `A` and `B` when your model is `lm(Y~A+B+A*B)`). When using Type II sums of squares, the ANOVA table will return the marginal effects of A and B *while igoring the interaction term*. However, when using Type III sums of squares, the effects of each predictor on our response variable will be returned AFTER the model accounts for the effects of all the other predictors. This means that (i) the order of predictors in your R code will not matter when using Type II or III and (ii) the results using Type II and III will be exactly equivalent *when you do not have an interaction term in the model*. We will use Type III sums of squares in this class, but here is an example illustrating the differences.

To use Type II or III sums of squares, install and load the `car` package, which has the function `Anova()` with a capital "A".

```
library(car)
```

Notice that both models now have the same exact ANOVA table. Bottom line: I recommend always using Type II or III sums of squares unless you have a very compelling reason to do otherwise.

When `f_YEAR` is listed first:

```
Anova(fit_ex1, type="III")
```

```
## Anova Table (Type III tests)
##
## Response: TICKS
##             Sum Sq  Df F value    Pr(>F)
## (Intercept)   7444   1  52.786 1.970e-12 ***
## f_YEAR        5450   2  19.321 9.788e-09 ***
## HEIGHT        6092   1  43.199 1.550e-10 ***
## Residuals    56268 399
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

When `HEIGHT` is listed first:

```
Anova(fit_ex2, type="III")
```

```
## Anova Table (Type III tests)
##
## Response: TICKS
##             Sum Sq  Df F value    Pr(>F)
## (Intercept)   7444   1  52.786 1.970e-12 ***
## HEIGHT        6092   1  43.199 1.550e-10 ***
## f_YEAR        5450   2  19.321 9.788e-09 ***
## Residuals    56268 399
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

Here are the example ANOVA tables for our three approaches. Read the vertical bar in "`HEIGHT|f_YEAR`" as "given", meaning we are looking at the effect of `HEIGHT` on `TICKS` given (or after accounting/adjusting for) the effcts of `f_YEAR` on `TICKS`.

Table 1: ANOVA Table for fit_ex1_df (sequential fit)

Predictor	DF	SS	MS	F	P
f_YEAR	2	7049.72	3524.86	24.99	<0.0001
HEIGHT|f_YEAR	1	6091.98	6091.98	43.2	<0.01
Residuals	399	56268.21	141.02		

Table 2: ANOVA Table for fit_ex2_df (sequential fit)

Predictor	DF	SS	MS	F	P
HEIGHT	1	7692.19	7692.19	54.55	<0.0001
f_YEAR|HEIGHT	2	5449.52	2724.76	19.32	<0.0001
Residuals	399	56268.21	141.02		

Table 3: ANOVA Table for fit_ex1_df (marginal fit)

Predictor	DF	SS	MS	F	P
f_YEAR|HEIGHT	2	5449.52	2724.7600	19.32	<0.0001
HEIGHT|f_YEAR	1	6091.98	6091.9800	43.2	<0.01
Residual	399	56268.21	141.0231		

6.3 A bit more on Type II vs. III Sums of Squares

Let's fit the same model as above (i.e., `lm(TICKS ~ f_YEAR + HEIGHT)`) but we will include an interaction term: `lm(TICKS ~ f_YEAR + HEIGHT + f_YEAR*HEIGHT`. Here is a reminder of that model:

Table 4: ANOVA Table for fit_ex1_df (marginal fit)

Predictor	DF	SS	MS	F	P
f_YEAR|HEIGHT	2	5449.52	2724.7600	19.32	<0.0001
HEIGHT|f_YEAR	1	6091.98	6091.9800	43.2	<0.01
Residual	399	56268.21	141.0231		

6.3.1 Type II

Note that in the model below, the estimates for the main effects (`f_YEAR, HEIGHT`) are the same as the "marginal" fit above in the model that did not have an interaction term (`fit_ex1_df`). There are slight differences due to rounding in the below output.

```
fit_ex_SS <- lm(TICKS ~ f_YEAR + HEIGHT + f_YEAR*HEIGHT, data=grouseticks)
Anova(fit_ex_SS, type="II")
```

```
## Anova Table (Type II tests)
##
## Response: TICKS
##               Sum Sq  Df F value    Pr(>F)
## f_YEAR          5450   2 20.1085 4.804e-09 ***
## HEIGHT          6092   1 44.9583 6.937e-11 ***
## f_YEAR:HEIGHT   2473   2  9.1271 0.0001332 ***
## Residuals      53795 397
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

6.3.2 Type III

Note that the below estimates for the main effects (`f_YEAR, HEIGHT`) are NOT the same as the "marginal" fit above in the model that did not have an interaction term (Table 4). The values associated with the interaction are the same, however: the below model provides the effects of `f_YEAR` and `HEIGHT` after accounting the effect of `f_YEAR*HEIGHT`.

```
Anova(fit_ex_SS, type="III")
```

```
## Anova Table (Type III tests)
##
## Response: TICKS
##               Sum Sq  Df F value    Pr(>F)
## (Intercept)     5915   1 43.6540 1.263e-10 ***
## f_YEAR          2911   2 10.7418 2.863e-05 ***
## HEIGHT          5210   1 38.4525 1.409e-09 ***
## f_YEAR:HEIGHT   2473   2  9.1271 0.0001332 ***
## Residuals      53795 397
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

7 R Activity

We will work with the same data set for the next two weeks. The data were generated from a study investigating the effects of animal species, diet, and drug type on blood glucose levels (mg/dl) using a $2 \times 3 \times 2$ factorial arrangement. Note that a factorial design of $A \times B \times C$ means that you designed a study looking at the effects of three factors, one having A levels, one having B levels, and one having C levels.

In this study, two animal species (goats or sheep) were fed one of three diets (control, alfalfa hay, and cottonseed meal) and received a drug injection (slaframine in saline or just saline). The 12 treatment combinations were assigned in a randomized complete block design with twelve blocks (replications). So, each combination of animal $\times$ diet $\times$ drug combination appears twelve times.

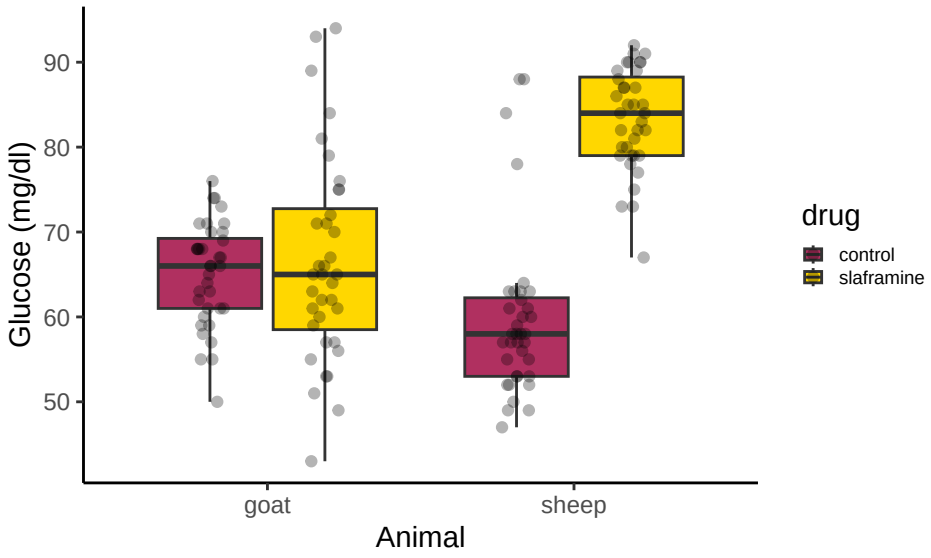
We are going to look at the effects of drug and animal on glucose blood levels and entirely ignore diet.

1. Load in the `glucose_df.txt` dataset.

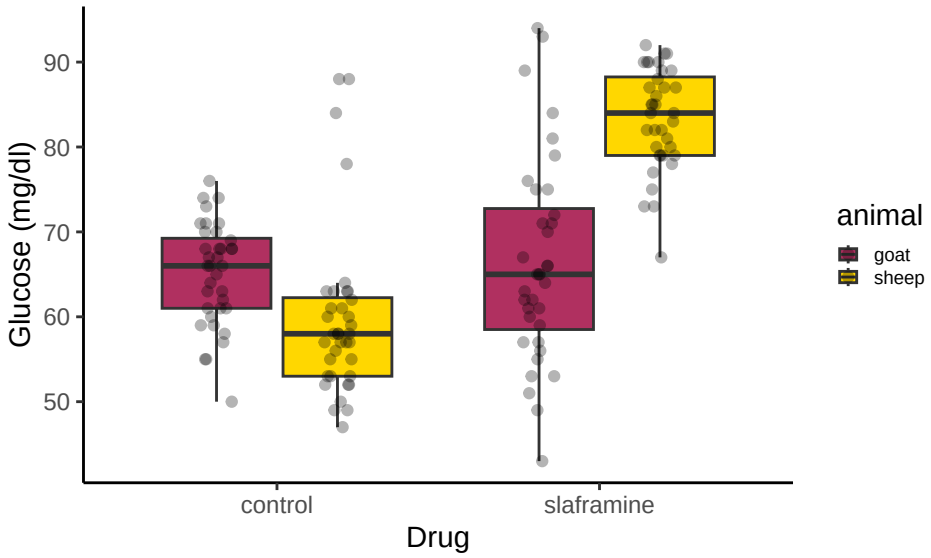
```
gluc_df <- read.table("glucose_df.txt", sep="\t", header=T)
summary(gluc_df)
```

```
##       rep           animal              diet               drug
##  Min.   : 1.00   Length:144         Length:144         Length:144
##  1st Qu.: 3.75   Class :character   Class :character   Class :character
##  Median : 6.50   Mode  :character   Mode  :character   Mode  :character
##  Mean   : 6.50
##  3rd Qu.: 9.25
##  Max.   :12.00
##     glucose
##  Min.   :43.00
##  1st Qu.:59.00
##  Median :66.00
##  Mean   :68.65
##  3rd Qu.:79.00
##  Max.   :94.00
```

2. Create a grouped boxplot of `glucose` as a function of `drug` and `animal` like the one below. Please change the colors to a combination of your choosing (Go Gophers :)) and make sure to overlay the raw data points on top of your boxes. Based on eyeballing the plot, provide a 1-2 sentence description of any pattern(s).




```
ggplot(gluc_df, mapping=aes( y=glucose, x=drug, fill=animal))+
geom_boxplot(outlier.color=NA)+
geom_point(position = position_jitterdodge( jitter.width = 0.1), alpha=0.3)+
theme_classic()+
ylab("Glucose (mg/dl)")+
xlab("Drug")+
theme(legend.text=element_text(size=7))+
theme(legend.key.size = unit(0.3, 'cm'))+
scale_fill_manual(values=c("maroon", "gold"))
```



ANSWER: It seems like glucose in the goats did not change with slaframine, whereas the glucose in sheeps increased with slaframine

3. We are interested in quantifying variation in glucose (our response variable). Note that we could analyze these data in the "historical" way by fitting `rep` (the column for blocks) as a so-called "fixed effect" (i.e., as a regular old predictor). Next week, you will get practice fitting mixed-effects models, in which `rep` would be fit as a so-called "random intercept" or "random effect". Do not worry, as I will also

further explain the terms fixed, random, and mixed-effect next week. However, for this week, we are going to simplify things: ignore `rep` and just fit `animal`, `drug`, and their interaction (`animal` × `drug`) as the predictors (again, we are ignoring the `diet` column).

```
fit_1 <- lm(glucose~drug*animal, data=gluc_df)
summary(fit_1)
```

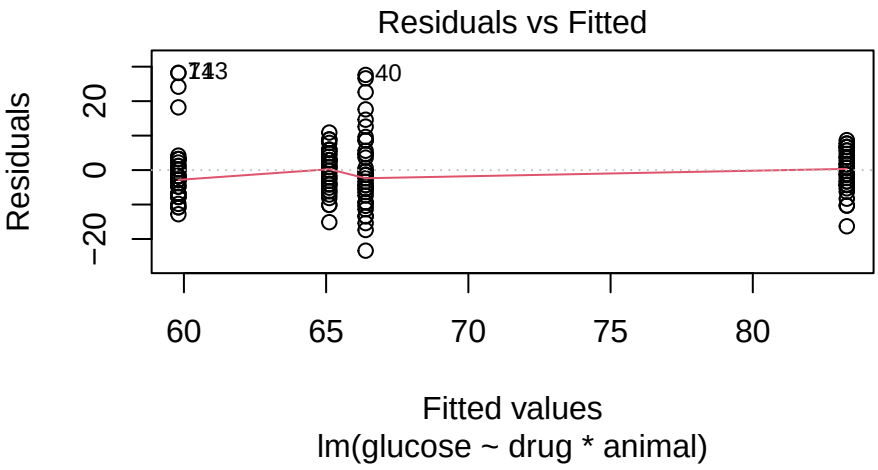
```
##
## Call:
## lm(formula = glucose ~ drug * animal, data = gluc_df)
##
## Residuals:
##      Min       1Q   Median       3Q      Max
## -23.3889  -5.1597  -0.9583   3.9653  28.1944
##
## Coefficients:
##                             Estimate Std. Error t value Pr(>|t|)
## (Intercept)                   65.111      1.483  43.919  < 2e-16 ***
## drugslaframine                 1.278      2.097   0.609   0.5432
## animalsheep                   -5.306      2.097  -2.531   0.0125 *
## drugslaframine:animalsheep    22.222      2.965   7.495 6.87e-12 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 8.895 on 140 degrees of freedom
## Multiple R-squared:  0.5024, Adjusted R-squared:  0.4917
## F-statistic: 47.11 on 3 and 140 DF,  p-value: < 2.2e-16
```

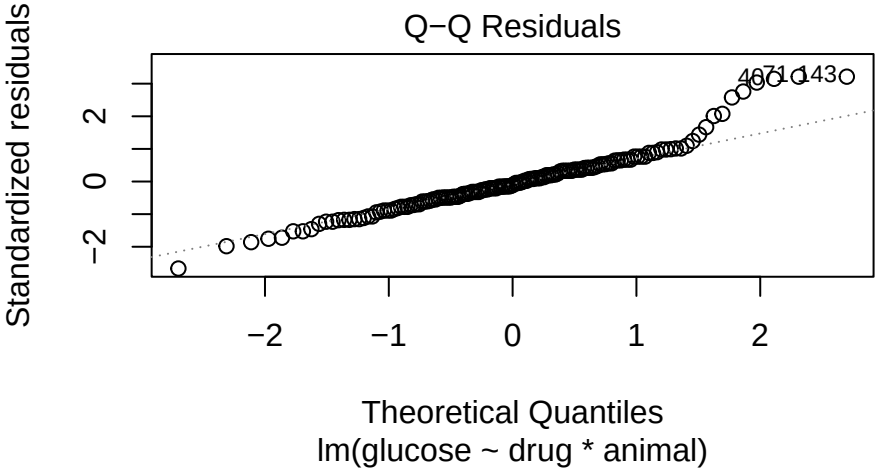
4. Run an `Anova()` on the model and ensure you are using marginal fits (Type III sums of squares).

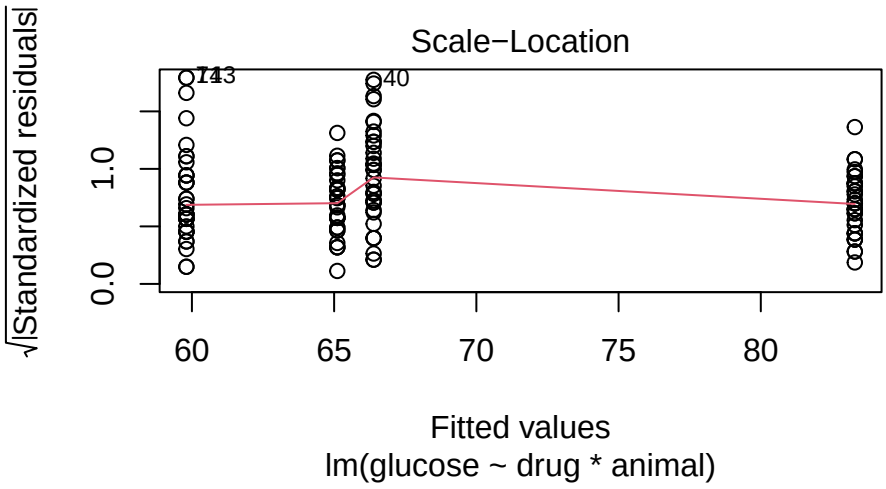
```
library(car)
Anova(fit_1, type="III")
```

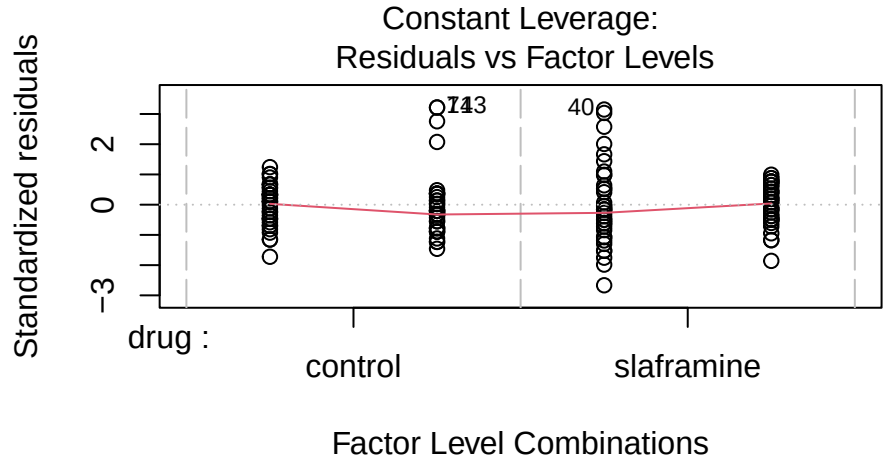
```
## Anova Table (Type III tests)
##
## Response: glucose
##             Sum Sq  Df   F value    Pr(>F)
## (Intercept) 152620   1 1928.8717 < 2.2e-16 ***
## drug            29   1    0.3714    0.5432
## animal         507   1    6.4036    0.0125 *
## drug:animal   4444   1   56.1705  6.87e-12 ***
## Residuals    11077 140
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

```
plot(fit_1)
```


Residuals vs Fitted
Residuals
−20
0
20
60
65
70
75
80
Fitted values
lm(glucose ~ drug * animal)


Q−Q Residuals
Standardized residuals
−2
0
2
−2
−1
0
1
2
Theoretical Quantiles
lm(glucose ~ drug * animal)





5. Conduct a pairwise comparisons of the interaction term

```
library(emmeans)
act_emm <- emmeans(fit_1, ~drug*animal)
pairs(act_emm)
```

```
##  contrast                          estimate  SE  df t.ratio p.value
##  control goat - slaframine goat       -1.28 2.1 140  -0.609  0.9289
##  control goat - control sheep          5.31 2.1 140   2.531  0.0596
##  control goat - slaframine sheep     -18.19 2.1 140  -8.678  <.0001
##  slaframine goat - control sheep       6.58 2.1 140   3.140  0.0110
##  slaframine goat - slaframine sheep  -16.92 2.1 140  -8.069  <.0001
##  control sheep - slaframine sheep    -23.50 2.1 140 -11.209  <.0001
##
## P value adjustment: tukey method for comparing a family of 4 estimates
```

6. Write 3-4 sentences interpreting the results of your analyses. Please try to write in biological terms, not statistical, but make sure to include the relevant summary statistics for any claims you make.

ANSWER: The response of glucose blood levels to slaframine varied by animal species (drug $\times$ animal interaction: $F_{1,140} = 56.17, p < 0.0001$). It appeared that sheep treated with slaframine had concentrations of glucose that were 23.5 units higher than control sheep ($t_{140} = 11.21, p < 0.0001$) and 16.9 units higher than goats treated with slaframine ($t_{140} = 8.07, p < 0.0001$). Additionally, there were no statistically clear differences in glucose levels between treated vs. control goats ($t_{140} = 0.61, p = 0.93$) and control sheep vs. control goats ($t_{140} = 2.53, p = 0.06$). Taken together, our findings seem to be driven by a positive response in glucose levels of sheep following slaframine treatment with no corresponding changes in glucose levels of goats.