load("dragons.RData")

head(dragons)

hist(dragons$testScore) # seems close to a normal distribution - good!

Chart, histogram

Description automatically generated

dragons$bodyLength2 <- scale(dragons$bodyLength, center = **TRUE**, scale = **TRUE**)

basic.lm <- lm(testScore ~ bodyLength2, data = dragons)

summary(basic.lm)

library(tidyverse) *# load the package containing both ggplot2 and dplyr*

(prelim\_plot <- ggplot(dragons, aes(x = bodyLength, y = testScore)) +

geom\_point() +

geom\_smooth(method = "lm"))

Chart, scatter chart

Description automatically generated

Plot the residuals: the red line should be nearly flat, like the dashed grey line:

plot(basic.lm, which = 1) *# not perfect...*

*## but since this is a fictional example we will go with it*

*## for your own data be careful:*

*## the bigger the sample size, the less of a trend you'd expect to see*

Chart, scatter chart

Description automatically generated

plot(basic.lm, which = 2) *# a bit off at the extremes, but that's often the case; again doesn't look too bad*

Chart, line chart

Description automatically generated

We collected multiple samples from eight mountain ranges. It’s perfectly plausible that the data from within each mountain range are more similar to each other than the data from different mountain ranges: they are correlated.

Have a look at the data to see if above is true:

boxplot(testScore ~ mountainRange, data = dragons) *# certainly looks like something is going on here*

Chart, box and whisker chart

Description automatically generated

(colour\_plot <- ggplot(dragons, aes(x = bodyLength, y = testScore, colour = mountainRange)) +

geom\_point(size = 2) +

theme\_classic() +

theme(legend.position = "none"))

Chart, scatter chart

Description automatically generated

From the above plots, it looks like our mountain ranges vary both in the dragon body length **AND** in their test scores. This confirms that our observations from within each of the ranges **aren’t independent**. We can’t ignore that: as we’re starting to see, it could lead to a completely erroneous conclusion.

We could run many separate analyses and fit a regression for each of the mountain ranges.

Lets have a quick look at the data split by mountain range. We use the facet\_wrap to do that:

(split\_plot <- ggplot(aes(bodyLength, testScore), data = dragons) +

geom\_point() +

facet\_wrap(~ mountainRange) + *# create a facet for each mountain range*

xlab("length") +

ylab("test score"))

Chart, scatter chart

Description automatically generated

mountain.lm <- lm(testScore ~ bodyLength2 + mountainRange, data = dragons)

summary(mountain.lm)

Now body length is not significant. But let’s think about what we are doing here for a second. The above model is estimating the difference in test scores between the mountain ranges - we can see all of them in the model output returned by summary(). But we are not interested in quantifying test scores for each specific mountain range: we just want to know whether body length affects test scores and we want to simply **control for the variation** coming from mountain ranges.

Note that **our question changes slightly here**: while we still want to know whether there is an association between dragon’s body length and the test score, we want to know if that association exists **after** controlling for the variation in mountain ranges.

We will fit the random effect using the syntax (1|variableName):

mixed.lmer <- lmer(testScore ~ bodyLength2 + (1|mountainRange), data = dragons)

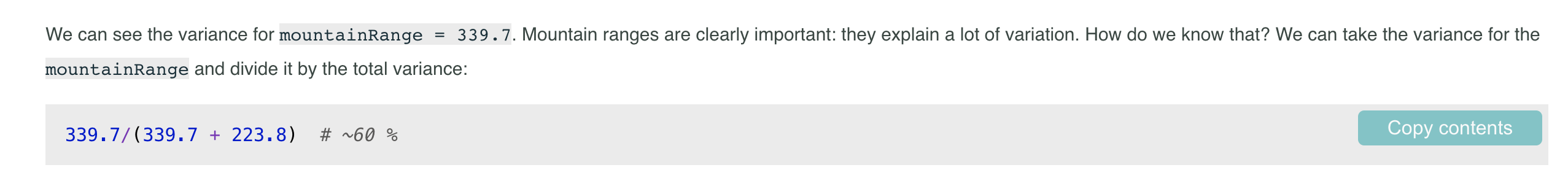
summary(mixed.lmer)

Once we account for the mountain ranges, it’s obvious that dragon body length doesn’t actually explain the differences in the test scores. *How is it obvious?* I hear you say?

Take a look at the summary output: notice how the **model estimate** is smaller than its associated error? That means that the effect, or slope, cannot be distinguised from zero.

Text

Description automatically generated



plot(mixed.lmer) *# looks alright, no patterns evident*

A picture containing text, sky

Description automatically generated

qqnorm(resid(mixed.lmer))

qqline(resid(mixed.lmer)) *# points fall nicely onto the line - good!*

**Chart, line chart

Description automatically generated**

just like we did with the mountain ranges, we have to assume that data collected within our sites might be **correlated** and so we should include sites as **an additional random effect** in our model.

Our site variable is a three-level factor, with sites called a, b and c. The nesting of the site within the mountain range is **implicit** - our sites are meaningless without being assigned to specific mountain ranges, i.e. there is nothing linking site b of the Bavarian mountain range with site b of the Central mountain range. To avoid future confusion we should create a new variable that is **explicitly nested**. Let’s call it sample:

dragons <- within(dragons, sample <- factor(mountainRange:site))

Our question gets **adjusted slightly again**: Is there an association between body length and intelligence in dragons **after** controlling for variation in mountain ranges and sites within mountain ranges?

mixed.lmer2 <- lmer(testScore ~ bodyLength2 + (1|mountainRange) + (1|sample), data = dragons) *# the syntax stays the same, but now the nesting is taken into account*

summary(mixed.lmer2)

Let’s plot this again - visualising what’s going on is always helpful. You should be able to see eight mountain ranges with three sites (different colour points) within them, with a line fitted through each site.

(mm\_plot <- ggplot(dragons, aes(x = bodyLength, y = testScore, colour = site)) +

facet\_wrap(~mountainRange, nrow=2) + *# a panel for each mountain range*

geom\_point(alpha = 0.5) +

theme\_classic() +

geom\_line(data = cbind(dragons, pred = predict(mixed.lmer2)), aes(y = pred), size = 1) + *# adding predicted line from mixed model*

theme(legend.position = "none",

panel.spacing = unit(2, "lines")) *# adding space between panels*

)

**Diagram

Description automatically generated with low confidence**