Biostatistics Task 3

Danyu Zhang & Daniel Alonso

May 28th, 2021

Exercise 1

Fitting a model with random effects

We fit the model with random effects as follows:

```
model <- lme(response~Treatment, random=list(~1|Block), data=pbib)
```

Is the type of fertilizer significant?

According to the model summary, no p-value is below our 0.05 significance level, therefore the treatment (type of fertilizer) cannot be considered significant

```
#> Linear mixed-effects model fit by REML
#>   Data: pbib
#>        AIC      BIC    logLik
#>   85.98489 116.6982 -25.99245
#>
#> Random effects:
#>  Formula: ~1 | Block
#>         (Intercept) Residual
#> StdDev:   0.2156896 0.292505
#>
#> Fixed effects:  response ~ Treatment
#>                  Value Std.Error DF   t-value p-value
#> (Intercept)  2.8175225 0.1664127 31 16.930935  0.0000
#> Treatment10 -0.3264609 0.2220608 31 -1.470142  0.1516
#> Treatment11  0.0811765 0.2272002 31  0.357291  0.7233
#> Treatment12  0.2352989 0.2220608 31  1.059615  0.2975
#> Treatment13 -0.1997533 0.2220608 31 -0.899543  0.3753
#> Treatment14 -0.3262114 0.2220608 31 -1.469018  0.1519
#> Treatment15  0.0417105 0.2220608 31  0.187834  0.8522
#> Treatment2  -0.4122076 0.2220608 31 -1.856282  0.0729
#> Treatment3  -0.3625793 0.2220608 31 -1.632793  0.1126
#> Treatment4  -0.0336921 0.2220608 31 -0.151725  0.8804
#> Treatment5  -0.0126245 0.2220608 31 -0.056852  0.9550
#> Treatment6   0.0931714 0.2272002 31  0.410085  0.6846
#> Treatment7  -0.0285375 0.2220608 31 -0.128512  0.8986
#> Treatment8  -0.0359170 0.2220608 31 -0.161744  0.8726
#> Treatment9   0.0737886 0.2220608 31  0.332290  0.7419
#>  Correlation:
#>             (Intr) Trtm10 Trtm11 Trtm12 Trtm13 Trtm14 Trtm15 Trtmn2 Trtmn3
#> Treatment10 -0.667
#> Treatment11 -0.683  0.512
#> Treatment12 -0.667  0.500  0.512
#> Treatment13 -0.667  0.500  0.512  0.500
#> Treatment14 -0.667  0.500  0.512  0.500  0.500
#> Treatment15 -0.667  0.477  0.512  0.500  0.500  0.500
#> Treatment2  -0.667  0.500  0.512  0.477  0.500  0.500  0.500
#> Treatment3  -0.667  0.500  0.512  0.500  0.477  0.500  0.500  0.500
#> Treatment4  -0.667  0.500  0.512  0.500  0.500  0.477  0.500  0.500  0.500
#> Treatment5  -0.667  0.477  0.512  0.500  0.500  0.500  0.477  0.500  0.500
#> Treatment6  -0.683  0.512  0.500  0.512  0.512  0.512  0.512  0.512  0.512
#> Treatment7  -0.667  0.500  0.512  0.477  0.500  0.500  0.500  0.477  0.500
#> Treatment8  -0.667  0.500  0.512  0.500  0.477  0.500  0.500  0.500  0.477
#> Treatment9  -0.667  0.500  0.512  0.500  0.500  0.477  0.500  0.500  0.500
#>             Trtmn4 Trtmn5 Trtmn6 Trtmn7 Trtmn8
#> Treatment10
#> Treatment11
#> Treatment12
#> Treatment13
#> Treatment14
#> Treatment15
#> Treatment2
#> Treatment3
#> Treatment4
#> Treatment5   0.500
#> Treatment6   0.512  0.512
#> Treatment7   0.500  0.500  0.512
#> Treatment8   0.500  0.500  0.512  0.500
#> Treatment9   0.477  0.500  0.512  0.500  0.500
#>
#> Standardized Within-Group Residuals:
#>         Min          Q1         Med          Q3         Max
#> -1.52465492 -0.45435363 -0.01326845  0.42567512  2.74433134
#>
#> Number of Observations: 60
#> Number of Groups: 15
```

What is the percentage of variability explained by the block effect?

We can see that the percentage of variability explained by the block effect is not very high:

```
#> [1] 0.2760676
```

Improving the model

We can make the model a little better by considering more random effects:

```r
model <- lme(response~Treatment, random=list(~1|Block, ~1|Treatment), data=pbib)
```

As this increases the explained variability of the model significantly (both variability explained by fixed and random effects):

```
#> [1] 0.9272038
```

Exercise 2

Identify random and fixed categorical variables

- **Thickness**: It's a continuous variable. For the purposes of this experiment, we will use it as the response variable.
- **Source**: We should consider it as a fixed effect. It might not be as advisable to use a variable with such little levels as a random effect.
- **Site**: As before, given the low amount of levels, it's also not advisable to use as a random effect. However, less than so for Source, it is less interesting for the experiment to see the difference between sites.
- **Lot**: We will treat this variable as a random effect. We suspect this could definitely represent a large portion of the variability. It also is categorical and has a lot of levels, therefore it's appropriate to consider it as such.
- **Wafer**: We treat it as a fixed effect, treating it as a random effect yields a programmatic error.

Modelling

We run the model using the *lmer* function, using the formula:

Thickness ~ **Source** + **Site** + **Wafer**

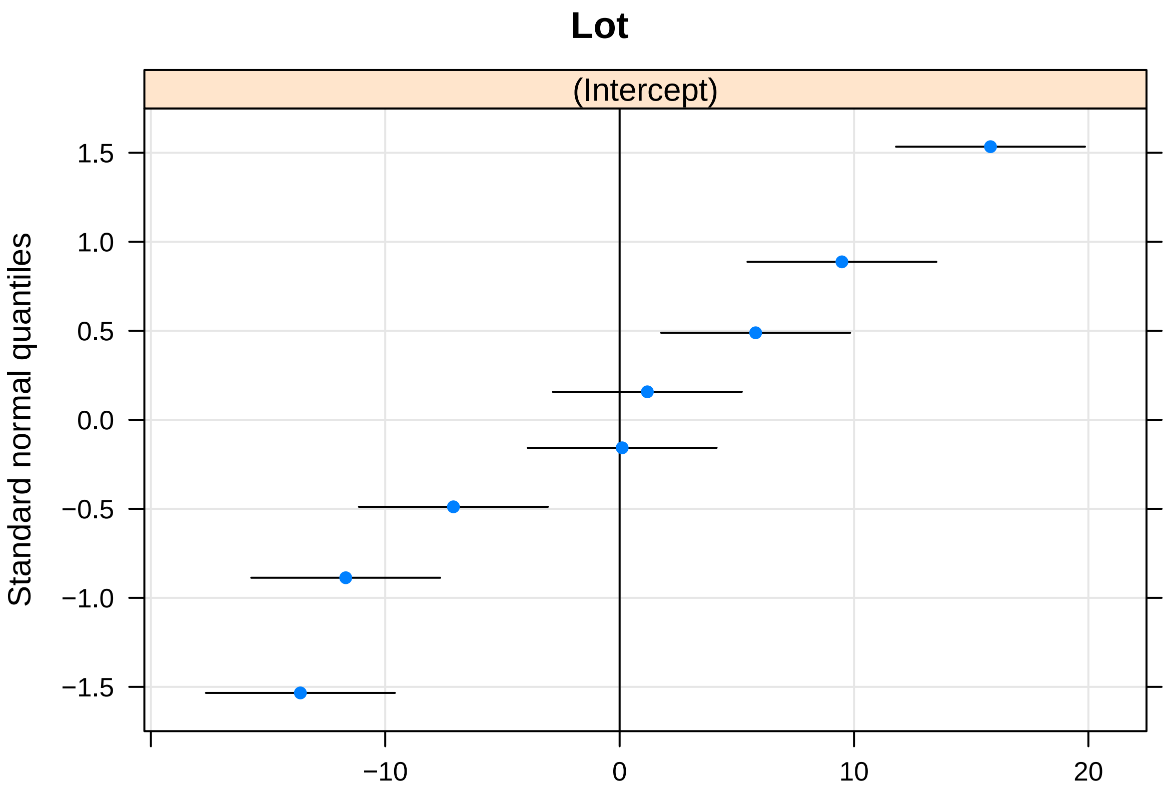
And considering the ~**1|Lot** random effect.

And we get the model as follows:

```r
model <- lmer(Thickness~Source + Site + Wafer + (1|Lot), data=oxide)
```

Analysis

One of our model assumptions is normality, as we can see, the lot random effect seems to follow a normal distribution.

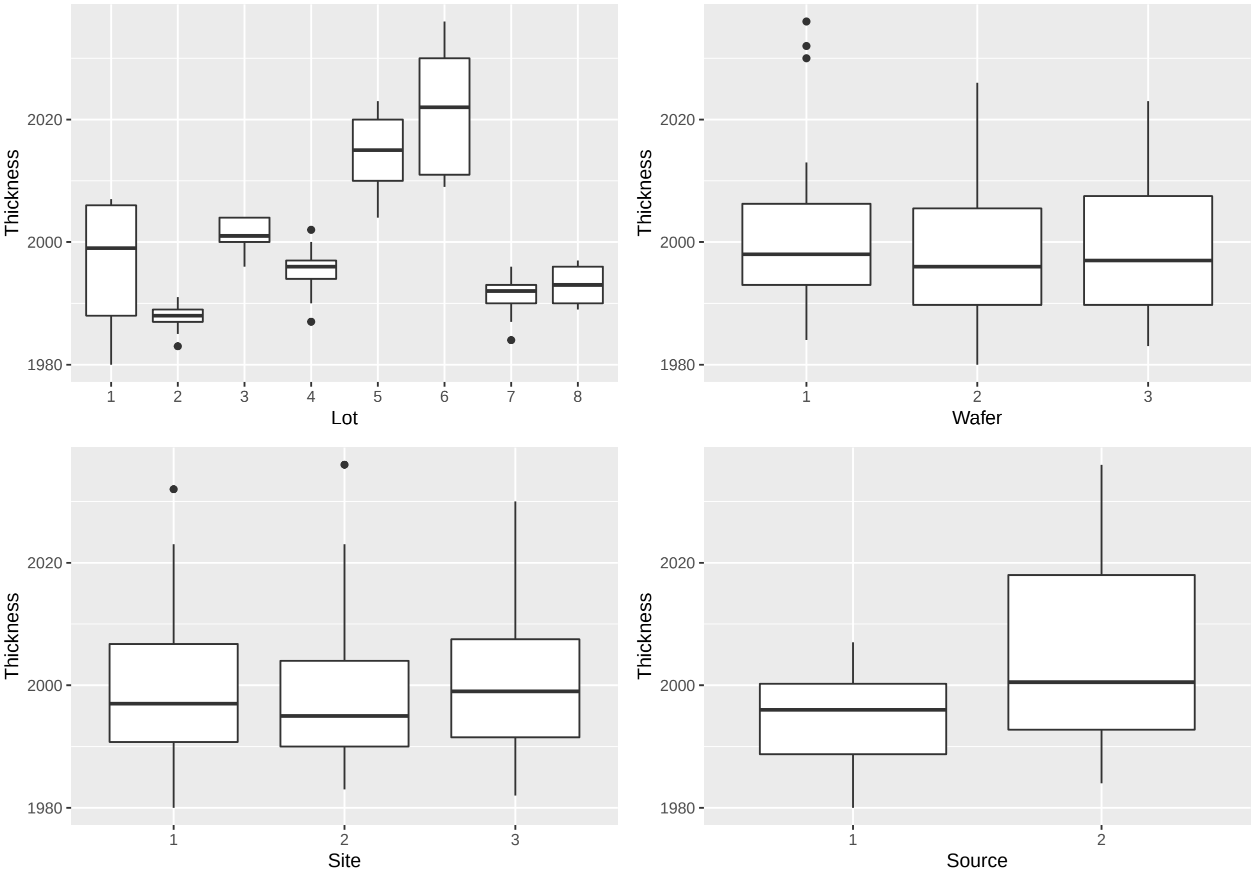


Looking at the model summary:

```
#> Linear mixed model fit by REML ['lmerMod']
#> Formula: Thickness ~ Source + Site + Wafer + (1 | Lot)
#>    Data: oxide
#> 
#> REML criterion at convergence: 467.9
#> 
#> Scaled residuals: 
#>      Min       1Q   Median       3Q      Max 
#> -2.47497 -0.43119  0.07816  0.49099  2.14419 
#> 
#> Random effects:
#>  Groups   Name        Variance Std.Dev.
#>  Lot      (Intercept) 128.87   11.352  
#>  Residual              39.37    6.275  
#> Number of obs: 72, groups:  Lot, 8
#> 
#> Fixed effects:
#>              Estimate Std. Error t value
#> (Intercept) 1996.8889     5.9580 335.158
#> Source2       10.0833     8.1622   1.235
#> Site2         -0.2500     1.8113  -0.138
#> Site3          0.8333     1.8113   0.460
#> Wafer2        -2.5417     1.8113  -1.403
#> Wafer3        -3.3750     1.8113  -1.863
#> 
#> Correlation of Fixed Effects:
#>         (Intr) Sourc2 Site2  Site3  Wafer2
#> Source2 -0.685                            
#> Site2   -0.152  0.000                     
#> Site3   -0.152  0.000  0.500              
#> Wafer2  -0.152  0.000  0.000  0.000       
#> Wafer3  -0.152  0.000  0.000  0.000  0.500
```

We can notice a few things. First of all, the explained variance by the Lot random effect is significant, at ~128.87, therefore we know that Lot is solid in the model as a random effect.

Looking at boxplots showing the differences among groups of each categorical variable, we can also see how this makes sense:



It's clear that there is significantly more variability when considering the lot, vs the other variables. The rest of the categorical variables don't seem to showcase major differences among groups.

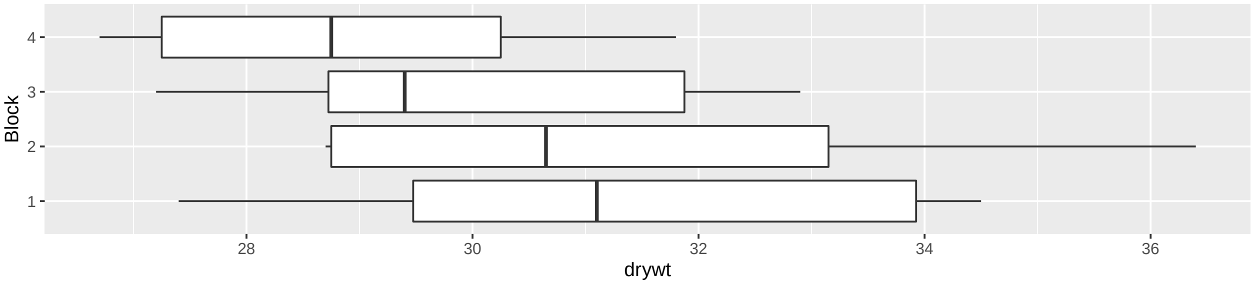
Exercise 3

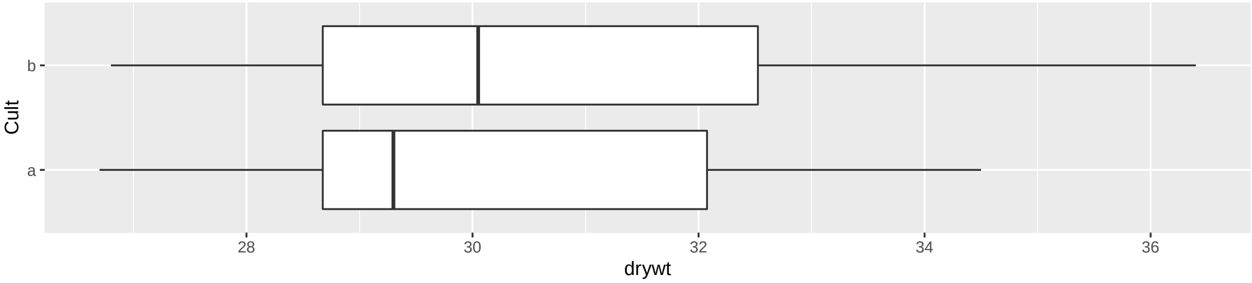
Modelling

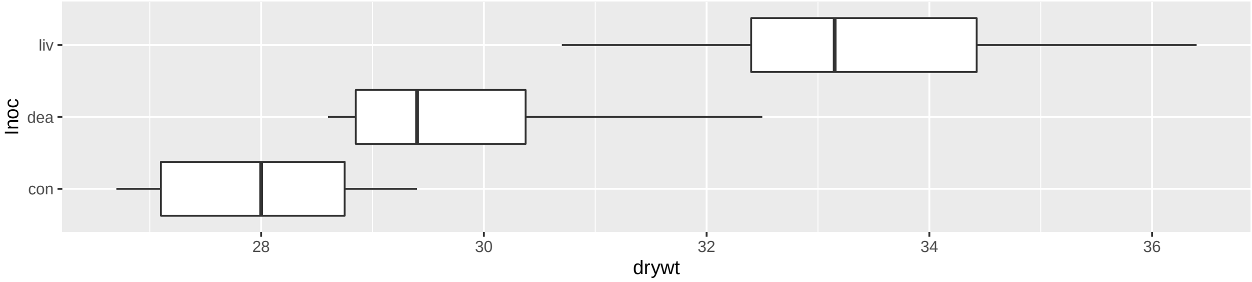
In order to check the variance explained by each variable as it is considered as random effect or not, we run the model and obtain the summary for each one of the models.

```
model1 <- lmer(drywt~Block + Cult + (1|Inoc), data=cultivation)
```

we find that the variable that most variance explains when ran as a random effect is *Inoc*, this finding, however, can be corroborated by visualizing a few boxplots viewing the data vs the categorical variables:







We can see that the largest accountability for differences can be given to *Inoc* followed by *Block*.

We could run the model considering these as nested random effects:

```
model2 <- lmer(drywt~Cult + (1|Inoc/Block), data=cultivation)
```

However, comparing the variability explained by the first vs the second model, we see that the simpler model, where we consider Block as a fixed effect is slightly better:

- Inoc as a sole random effect + block as fixed effect:

```
#>              R2m       R2c
#> [1,] 0.1254546 0.8789701
```

- Inoc/Block as a nested random effect

```
#>               R2m       R2c
#> [1,] 0.01109292 0.8386358
```

As a last resort, perhaps we could consider a model where both *Block* and *Inoc* are considered as random effects, but separately from one another:

```
model3 <- lmer(drywt~Cult + (1|Inoc) + (1|Block), data=cultivation)
```

And we can see there's a very minimal improvement of the model:

```
#>               R2m       R2c
#> [1,] 0.01076281 0.8803783
```

Therefore we would keep this as the final model, although for simplicity, if it were costlier to undergo in terms of materials/control of the environment for the study, there is zero harm in using the model where we only consider *Inoc* as a random effect. As the improvement is so minimal that we could say is nearly negligible.

Exercise 4

First we add the *math.8* and *math.11* average per school to the dataset:

```
maths$mean8_per_school <- sapply(maths$school, function(s) {mean(maths[maths$school == s,'math.8'])})
```

And then we calculate the difference between the students' *math.8* score and their respective school mean:

```
maths$math8_diff <- maths$math.8 - maths$mean8_per_school
```

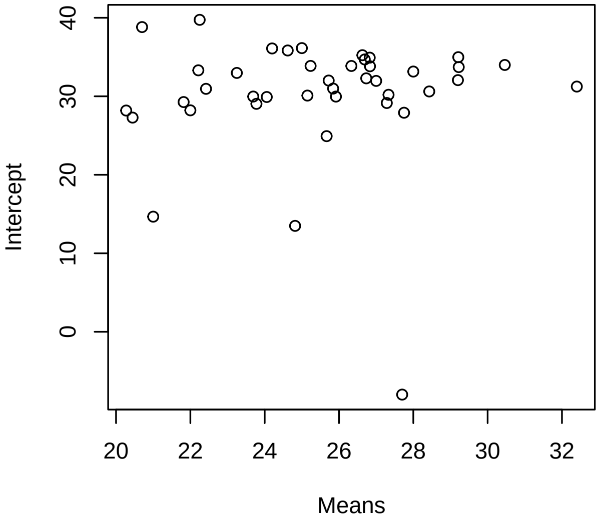
Fitting one model per school

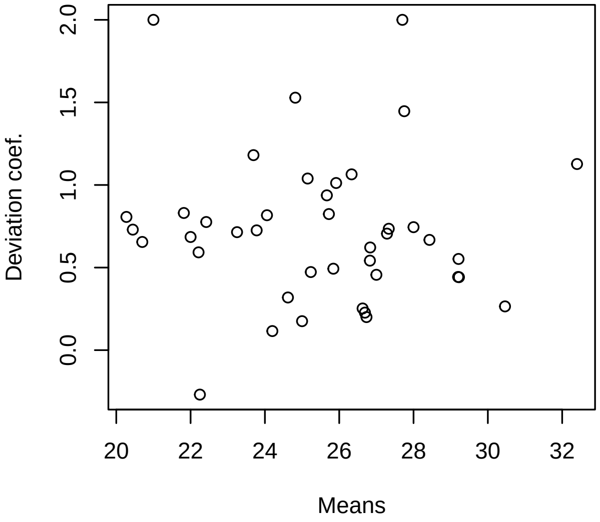
Fitting the model we obtain the following warning:

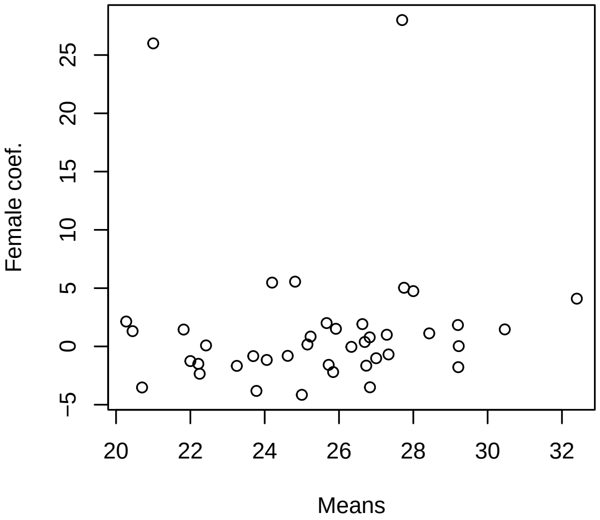
```
#> Warning in lmList(math.11 ~ math8_diff + female + manual | school, data = maths): Fitting failed for
#>   contrasts can be applied only to factors with 2 or more levels
```

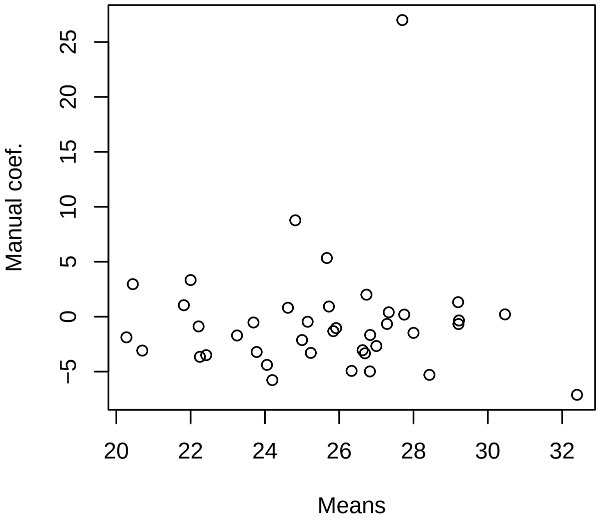
This tells us that there might be some of these schools where only one level of our chosen categorical variables is represented.

Plotting each set of coefficients









Checking the means vs teh coefficients we can notice a few things. First of all, they're all tightly packed around the same range, and the variability among the values isn't too high, with the exception with the deviation vs means plot.

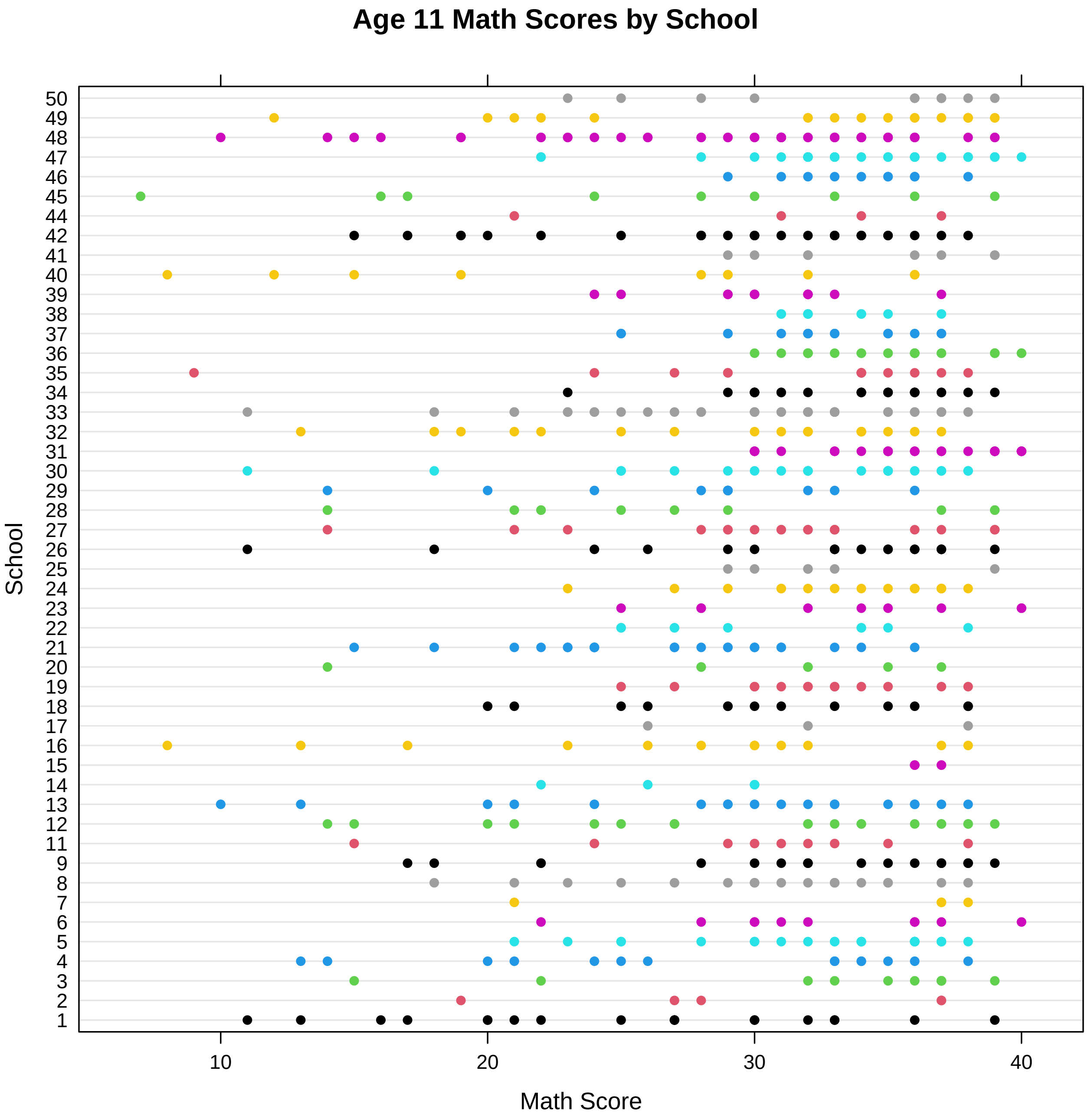
We can see which coefficients are also more significant than the others, and for example, sex doesn't seem to have a huge effect, at least visually, along with manual. However, given the scales, our perception might be erroneous.

In an effort to be precise, it would be reasonable to check the mean of each coefficient:

```
#> [1] 30.18311
#> [1] 0.7271331
#> [1] 1.447808
#> [1] -0.4582383
```

Surprisingly, sex seems to be the most meaningfully large coefficient, however, this might be due to the fact that it has significant outliers. If these were excluded, the result would've been noticeably different.

Looking at within-school variation for math.11 score



Here we notice a few things, first of all, in the grand scheme of things, there seems to not be a ton of differences among schools. Most scores cluster between 30 and 40, but there are clearly some schools with a significantly larger amount of data than others. Therefore there should be some schools which stand out, for example, school 15 only has scores in the upper-middle end of 30-40, but school 1 has a very even distribution of scores, so these should be considered differently, both in terms of results (difference in values) but also in amount of data.

So then we could say yes, the results might vary significantly, but also at the same time, we must be pragmatic about it and keep in mind that the difference in amount of data could be skewing our perspective.

Fitting a random effects model

Given that we're only interested in the within-school variation of these scores, we should formulate a model where only this is considered, however, prior to doing this, we will check a model including all the categorical variables as fixed effeects, in order to discard those not useful ones.

We see something interesting with respect to manual. As for the coefficients go though, we would say that whether the person is female or not would produce a change in 0.386 in the score, either positively (if female), or negatively (if male).

In the case of manual, there seems to be a significant value to its coefficient, where the score is reduced by 2.92 depending whether it's manual or not, if not, then it wouldn't be reduced by 2.92.

According to the p-val, manual should be significant, initially we thought this might be an error, but it seems about right.

```
#> Linear mixed-effects model fit by REML
#>   Data: maths
#>        AIC      BIC    logLik
#>   4753.977 4776.908 -2371.988
#>
#> Random effects:
#>  Formula: ~1 | school
#>         (Intercept) Residual
#> StdDev:    2.164125  6.10412
#>
#> Fixed effects:  math.11 ~ female + manual
#>                Value Std.Error  DF  t-value p-value
#> (Intercept) 32.49265 0.5863088 678 55.41901  0.0000
#> female1      0.38601 0.4661318 678  0.82812  0.4079
#> manual1     -2.92483 0.5153277 678 -5.67567  0.0000
#>  Correlation:
#>         (Intr) femal1
#> female1 -0.408
#> manual1 -0.598 -0.017
#>
#> Standardized Within-Group Residuals:
#>        Min         Q1        Med         Q3        Max
#> -3.7781802 -0.5023313  0.1950861  0.6783059  1.9230296
#>
#> Number of Observations: 728
#> Number of Groups: 48
```

Keeping manual and adding math8 scores

From the model summary we can see that both math.8 scores and manual remain significant, however, the introduction of this variable seems to reduce the significance of manual1, perhaps because the scores might be strongly correlated.

```
#> Linear mixed-effects model fit by REML
#>   Data: maths
#>        AIC     BIC   logLik
#>   4306.919 4329.85 -2148.46
#>
#> Random effects:
#>  Formula: ~1 | school
#>         (Intercept) Residual
#> StdDev:    1.823253 4.441594
#>
#> Fixed effects:  math.11 ~ manual + math.8
#>                 Value Std.Error  DF   t-value p-value
#> (Intercept) 14.701419 0.8269205 678 17.778515  0.0000
#> manual1     -0.712679 0.3869210 678 -1.841923  0.0659
#> math.8       0.638976 0.0254456 678 25.111427  0.0000
#>  Correlation:
#>         (Intr) manul1
#> manual1 -0.502
#> math.8  -0.866  0.226
#>
#> Standardized Within-Group Residuals:
#>         Min          Q1         Med          Q3         Max
#> -3.86047030 -0.50428777  0.07003699  0.58115688  3.33237257
#>
#> Number of Observations: 728
#> Number of Groups: 48
```

We can test for this correlation:

```
#> [1] 0.6838919
```

And we see a reasonable correlation, which for the purposes of this model, we can say is strong enough to be significant in a predictive model.

Testing whether to exclude the random effects or not

We test using the LRT:

```
#>               Model df      AIC      BIC    logLik   Test  L.Ratio p-value
#> model             1  5 4306.919 4329.850 -2148.459
#> reduced_model     2  4 4369.360 4387.705 -2180.680 1 vs 2 64.44143  <.0001
```

Given our p-val, which is minimal even for high significance levels (above 99%), we cannot reject that the random effect is not significant, therefore we must keep the full model which includes the random effects.

Assessing model quality

- R-squared for the model including all categorical variables:

```
#> [1] 0.1498581
```

- R-squared for the model including math.8, manual and the random effects

```
#>            R2m       R2c
#> [1,] 0.4724675 0.5485411
```

- R-squared for the model excluding the random effects

```
#>            R2m       R2c
#> [1,] 0.4692604 0.4692604
```

This shows, once again, that our random effects improve the model accuracy, and that math.8 is a significantly more valuable variable at the moment of creating the model.

All this can be confirmed, although the variability explained is still rather poor (<0.6).