Chapter 9: Linear models with both numeric and factor explanatory variables

Part 2: The model without interaction

STATS 201/8

University of Auckland

Learning Outcomes

In this chapter you will learn about:

- More about linear models with both categorical and numeric explanatory variables
- Model selection using Occam's razor
- The `anova` function for joint hypothesis tests
- Fitting a model without an interaction term
- Interpretting fitted models with factors having more than two levels
- Changing the reference level of a categorical variable
- Relevant `R`-code.

Section 9.1
Example: Using both IQ and teaching method to explain increase in language proficiency

Student language score by teaching method and IQ

In the following example, educational experts were interested in which of three different teaching methods was most effective in increasing a language score for students of a range of abilities – as measured by IQ.

In order to do this, 30 students were randomly allocated into three groups and each group was taught using a different teaching method. Each student's IQ was measured before the teaching programme began.

This randomisation was done to ensure that a range of student abilities was represented in each group. As students were in a test environment we can assume that their test scores are independent of each other.

The variables measured were:

`lang` the language test score achieved by the student after instruction
`IQ` the student's IQ
`method` the teaching method used (1, 2 or 3)

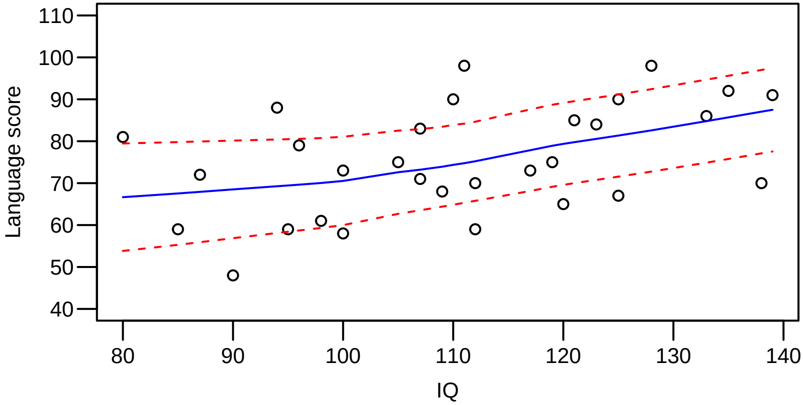
Student language score by teaching method and IQ...

It was of interest to see if there was any difference in the students' expected language score between teaching methods, and if this difference depended on IQ.

As usual, we begin by inspecting the data:

```
> ## Invoke the s20x library
> library(s20x)
> ## Importing data found in the s20x library into R
> data(teach.df)
> ## Plot the data with trendscatter()
> trendscatter(lang ~ IQ, f = 0.8, ylim = c(40, 110), data = teach.df,
+              ylab="Language score")
> ## Note that f is the proportion of points in the plot which influence the
> ## smooth at each value. Larger values of f give more smoothness!
```

Student language score by teaching method and IQ. . .



Hmmm – positive relationship with IQ suggested, but statistical significance not obvious. A plot that shows the teaching method would be more useful.

Student language score by teaching method and IQ...

In dataframe teach.df the method is recorded as a number, 1, 2 or 3.

```
> teach.df$method
 [1] 1 1 1 1 1 1 1 1 1 1 2 2 2 2 2 2 2 2 2 2 3 3 3 3 3 3 3 3 3 3
> class(teach.df$method)
[1] "integer"
```

However, these are just labels and could as easily have been ”A”, ”B” or ”C”. So, method needs to be ‘coerced’ into a factor so that it does not get treated as a numeric variable[1].

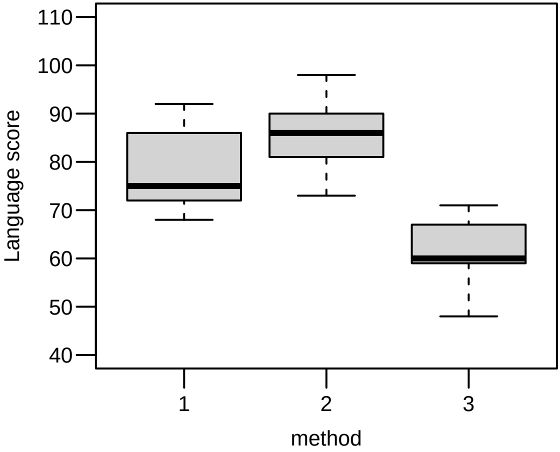
```
> teach.df$method = factor(teach.df$method)
> teach.df$method
 [1] 1 1 1 1 1 1 1 1 1 1 2 2 2 2 2 2 2 2 2 2 3 3 3 3 3 3 3 3 3 3
Levels: 1 2 3
> class(teach.df$method)
[1] "factor"
```

Not much has changed, but when we use method as the explanatory variable, plot functions and lm will now do the right thing since it is now a factor variable.

[1] What would happen if lm was treated method as numeric???

Student language score by teaching method and IQ...

```
> plot(lang ~ method, ylim=c(40,110), data=teach.df, ylab="Language score")
```

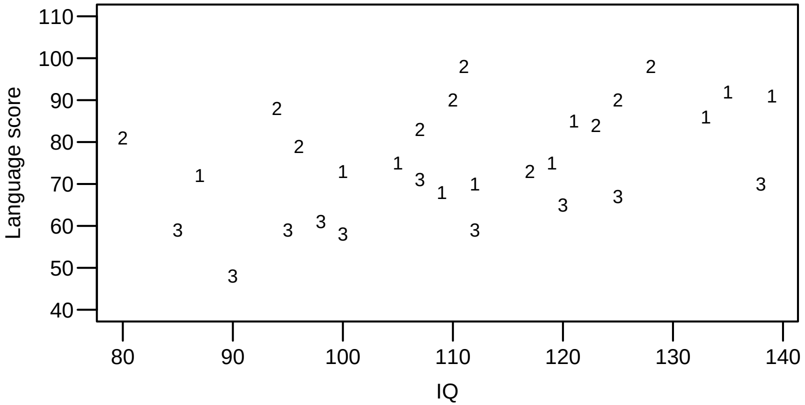


Teaching method 2 seems best. The worst appears to be method 3.

Student language score by teaching method and IQ. . .

A more useful plot:

```
> plot(lang ~ IQ, ylim=c(40,110), pch=as.character(method), data = teach.df,
+       ylab="Language score")
```



What do you see in this plot?

Student language score by teaching method and IQ...

What we see here is that the data for methods 1, 2, 3 seem to be scattered around approximately parallel straight lines. This means the `IQ` effect appears the same regardless of the teaching method. Conversely, the `Method` effect appears to be the same regardless of IQ.

In other words, it looks like `IQ` and `method` **do not** interact, i.e., there is no need to complicate the model by using a different slope for each method.

When we look at this data we see that the method 2 students are (on average) doing considerably better than method 1 students, who are (on average) better than the method 3 students.

So, we suspect that we there may be a difference in the expected teaching scores between methods, and that the effect of IQ may be the same regardless of method. We'll need to fit an appropriate linear model to find out for sure.

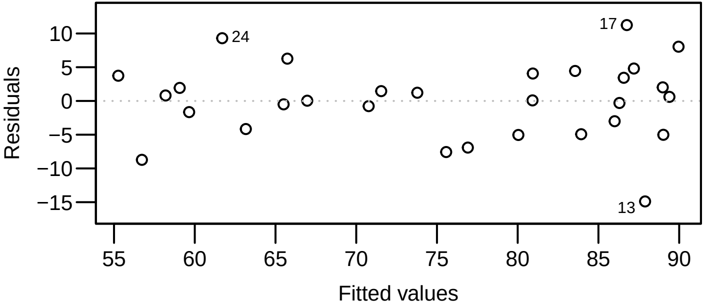
Student language score by teaching method and IQ. . .

Fitting and checking the interaction model

We will fit the model with interaction first, anticipating that the interaction will not be significant.

Here goes. . . , along with the usual assumption checks.

```
> TeachIQmethod.fit=lm(lang~IQ*method, data=teach.df)
> plot(TeachIQmethod.fit, which = 1)
```

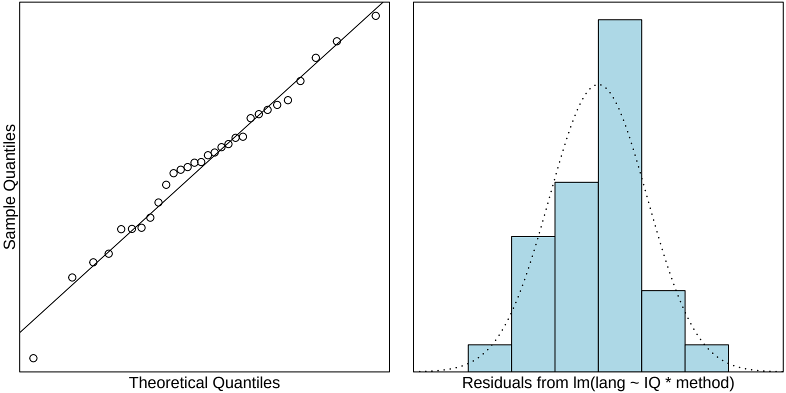


The **EOV** and no-trend assumptions seem to be okay.

Student language score by teaching method and IQ...

Fitting and checking the interaction model...

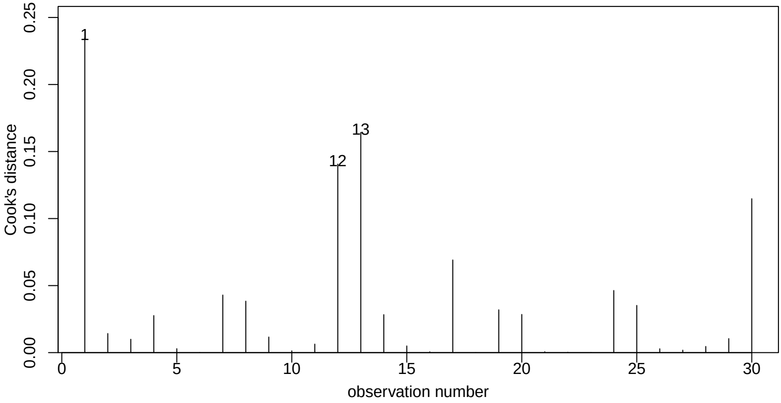
```
> normcheck(TeachIQmethod.fit)
```



Student language score by teaching method and IQ. . .

Fitting and checking the interaction model. . .

```
> cooks20x(TeachIQmethod.fit)
```



It looks like we can trust the output of the fitted model.

Section 9.2
Model selection using Occam's razor

Student language score by teaching method and IQ...

The fitted interaction model

Our fitted interaction model is:

```
> summary(TeachIQmethod.fit)
```

```
Coefficients:
            Estimate Std. Error t value Pr(>|t|)
(Intercept)  26.8346    14.5250   1.847  0.07704 .
IQ            0.4471     0.1241   3.604  0.00142 **
method2      39.0098    20.7473   1.880  0.07227 .
method3       3.5617    19.7222   0.181  0.85820
IQ:method2   -0.2587     0.1831  -1.413  0.17042
IQ:method3   -0.1546     0.1749  -0.883  0.38574
---
Residual standard error: 6.199 on 24 degrees of freedom
Multiple R-squared:  0.8121,Adjusted R-squared:  0.7729
F-statistic: 20.74 on 5 and 24 DF,  p-value: 5.284e-08
```

Model selection using Occam's razor

In previous Chapters we've seen that we remove terms that are not significant if doing so simplifies the fitted model. This is a very important principle of model selection and is an application of the "principle of parsimony", also known as **Occam's Razor**.[2] :

"The principle states that among competing models that predict equally well, the one with the fewest parameters should be selected."

In STATS20x we sometimes call it the **KISS** principle – "keep it simple, statistician".

In this class the general model selection approach we use is to do a hypothesis test to determine if we can remove the most complicated term from our current model.[3]

[2] Named after William of Ockham (c. 1287-1347), who was an English Franciscan friar and scholastic philosopher and theologian.

[3] In STATS 330 you will several holistic approaches to model selection that are based on the estimated predictive ability of the model.

Student language score by teaching method and IQ...

Model selection using Occam's razor

In our above interaction model for language score, the interaction term is the most complicated term, so we need to test whether the interaction is significant.

Hmmm, the null hypothesis of no interaction is saying that the last two coefficients in the summary table are both zero. That is, $H_0 : \beta_4 = \beta_5 = 0$.[4]

To do a single test about whether two or more coefficients are zero we need to produce an ANOVA table, using the `anova` function.

[4] A hypothesis that specifies the value of two or more coefficients is called a **joint** hypothesis.

Student language score by teaching method and IQ...

Model selection using Occam's razor...

```
> anova(TeachIQmethod.fit)
Analysis of Variance Table

Response: lang
          Df  Sum Sq Mean Sq F value    Pr(>F)
IQ         1 1004.42 1004.42 26.1416 3.124e-05 ***
method     2 2901.83 1450.91 37.7625 3.867e-08 ***
IQ:method  2   78.82   39.41  1.0257    0.3737
Residuals 24  922.13   38.42
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

The *P*-value associated with the interaction term `IQ:method` is large. So we conclude we do not have evidence of an interaction.

This means we do not need to compute a different slope for each teaching method – just different intercepts. Our instincts were right. This saves us having to have an extra two parameters in our final model, and simplifies the interpretation of our model since the lines are parallel.

Understanding the ANOVA table

An `anova` table is based on partitioning the total sums of squares, TSS. Recall, this is the residual sums of squares of the null model. The TSS is given by summing the values in the `Sum sq` column of the table.

In the above table, TSS $= 1004.42 + 2901.83 + 78.82 + 922.13 = 4907.2$. The amount of unexplained variability in our fitted model is given by the `Residuals` value of 922.13. This means that $922.13/4907.2 = 0.188$ is the proportion of the total variability that is not explained. The R^2 is therefore $1 - 0.188 = 0.812$.

Now, back to our example, where we now need to consider the formulation of the parallel line model.

Student language score by teaching method and IQ...

Fitting the model without interaction

Occam's razor dictates that we sharpen our model by removing the interaction term. To do this, we simply replace * by + in the model formula.

Non-interaction models are sometimes referred to as *additive models* (as the effects 'add up'), or 'main effects' models.

```
> TeachIQmethod.fit2=lm(lang~IQ+method, data=teach.df)
> summary(TeachIQmethod.fit2)
```

```
Coefficients:
             Estimate Std. Error t value Pr(>|t|)
(Intercept)  42.08552    8.73921   4.816 5.47e-05 ***
IQ            0.31564    0.07341   4.299 0.000213 ***
method2       9.87793    2.82068   3.502 0.001688 **
method3     -14.15922    2.85240  -4.964 3.70e-05 ***
---
Residual standard error: 6.205 on 26 degrees of freedom
Multiple R-squared:  0.796,Adjusted R-squared:  0.7725
F-statistic: 33.82 on 3 and 26 DF,  p-value: 3.986e-09
```

All the assumptions are fine (not shown) – not surprising, as all we have done is remove a term that was not significant.

Student language score by teaching method and IQ...

Interpretting the no-interaction model

The equation for the parallel lines (i.e, no-interaction) model is

$$\texttt{lang} = \beta_0 + \beta_1 \times \texttt{IQ} + \beta_2 \times \texttt{D2} + \beta_3 \times \texttt{D3} + \varepsilon$$

where, as usual $\varepsilon \overset{iid}{\sim} N(0, \sigma^2)$.

There are two indicator variables since teaching method has three levels:

- `D2` is an indicator variable whereby: `D2 = 1` if teaching method 2 is taught – otherwise it is 0.
- `D3` is an indicator variable whereby: `D3 = 1` if teaching method 3 is taught – otherwise it is 0.
- Teaching method 1 is the **reference/baseline** level group.

β_2 and β_3 represent the change in expected score (for any fixed student IQ) when we compare teaching method 2 or 3 to teaching method 1 (baseline).

Student language score by teaching method and IQ...

Model selection using Occam's razor...

The KISS principle requires us to determine if we can further simplify our model.

The no-interaction model can be simplified by removing the `IQ` term and/or the `method` terms. We see immediately from the summary table that `IQ` is highly significant, and so should not be removed.

To see if the `method` terms can be removed we want to test the joint null hypothesis that the intercepts are all identical, $H_0 : \beta_2 = \beta_3 = 0$. This would simplify our model to a simple linear regression model.

Recall: To do a test about whether two or more coefficients are zero we need to produce an ANOVA table, using the `anova` function.

Student language score by teaching method and IQ. . .

Model selection using Occam’s razor. . .

Let us see if we really do have identical intercepts.

```
> anova(TeachIQmethod.fit2)
Analysis of Variance Table

Response: lang
          Df Sum Sq Mean Sq F value    Pr(>F)
IQ         1 1004.4  1004.4  26.090 2.529e-05 ***
method     2 2901.8  1450.9  37.688 2.077e-08 ***
Residuals 26 1001.0    38.5
---
Signif. codes:  0 ’***’ 0.001 ’**’ 0.01 ’*’ 0.05 ’.’ 0.1 ’ ’ 1
```

The *P*-value associated with the `method` term is very small, so we conclude that the intercepts are different.

We do have to fit different intercepts for each teaching method. Our instincts were right.

Student language score by teaching method and IQ. . .

Our preferred model

Our preferred model is the no-interaction model:

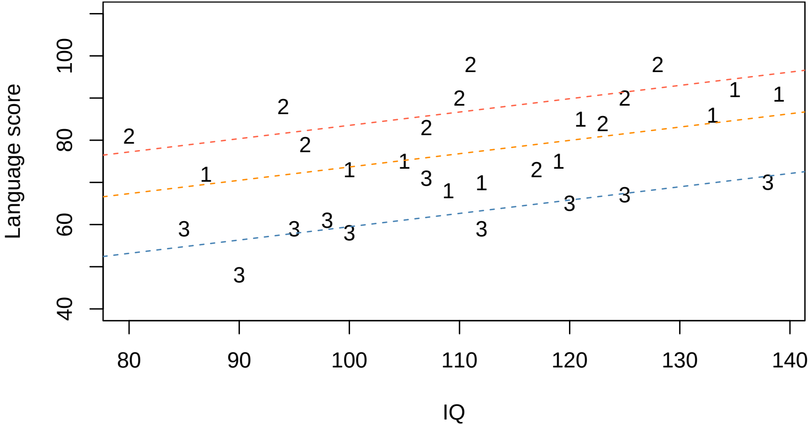
```
> summary(TeachIQmethod.fit2)
```

```
Coefficients:
             Estimate Std. Error t value Pr(>|t|)
(Intercept)  42.08552    8.73921   4.816 5.47e-05 ***
IQ            0.31564    0.07341   4.299 0.000213 ***
method2       9.87793    2.82068   3.502 0.001688 **
method3     -14.15922    2.85240  -4.964 3.70e-05 ***
---
Residual standard error: 6.205 on 26 degrees of freedom
Multiple R-squared:  0.796,Adjusted R-squared:  0.7725
F-statistic: 33.82 on 3 and 26 DF,  p-value: 3.986e-09
```

Student language score by teaching method and IQ...

The fitted model

```
> plot(lang ~ IQ, ylim=c(40,110), pch=as.character(method), data = teach.df,
+       ylab="Language score")
> b = coef(TeachIQmethod.fit2)
> abline(b[1], b[2], lty = 2, col = "darkorange")
> abline(b[1] + b[3], b[2], lty = 2, col = "tomato")
> abline(b[1] + b[4], b[2], lty = 2, col = "steelblue")
```



Student language score by teaching method and IQ…

Interpreting the output…

We are now able to deduce:

- $\beta_1 > 0 \implies$[5] IQ has a common positive effect on the expected language score of all students
- $\beta_2 > 0 \implies$ teaching method 2 is better than teaching method 1 regardless of a student's IQ.
- $\beta_3 < 0 \implies$ teaching method 3 is worse than teaching method 1 regardless of a student's IQ.

Our initial hunch about the best model has been justified.

[5] $\implies$ means *implies that*.

Section 9.3
Changing the reference level of teaching method

Student language score by teaching method and IQ. . .

Changing the reference level

The above statements are useful, but are incomplete – we still don't know how method 2 differs from method 3?[6]

The lm function has chosen the baseline (i.e., reference) level to be method 1 since "1" comes before "2" and "3" when these values are sorted by a computer. We need to change this to make method 2 (or alternatively method 3) the baseline. The fitted model will be exactly the same, but the intercept coefficients will change due to the change in reference level.

Here is the R code to do this:

```
> teach.df$method = relevel(teach.df$method, ref = "2")
> TeachIQmethod.fit3=lm(lang~IQ+method, data=teach.df)
```

Let us compare the summarys of TeachIQmethod.fit3 and TeachIQmethod.fit2.

[6] The issue of estimating all pairwise differences between factor levels is an issue we deal with more fully in a later chapter.

Student language score by teaching method and IQ...

Interpreting the output

```
> summary(TeachIQmethod.fit3)
```

```
Coefficients:
             Estimate Std. Error t value Pr(>|t|)
(Intercept)  51.96345    8.24637   6.301 1.14e-06 ***
IQ            0.31564    0.07341   4.299 0.000213 ***
method1      -9.87793    2.82068  -3.502 0.001688 **
method3     -24.03715    2.77910  -8.649 3.97e-09 ***
---
```

```
> summary(TeachIQmethod.fit2)
```

```
Coefficients:
             Estimate Std. Error t value Pr(>|t|)
(Intercept)  42.08552    8.73921   4.816 5.47e-05 ***
IQ            0.31564    0.07341   4.299 0.000213 ***
method2       9.87793    2.82068   3.502 0.001688 **
method3     -14.15922    2.85240  -4.964 3.70e-05 ***
---
```

Student language score by teaching method and IQ...

Interpreting the output...

Note that the `method1` coefficient in `fit3` has the same magnitude, but opposite sign to `method2` in `fit2`. This is because `fit3` measures how method 1 differs form method 2 whereas `fit2` measures how method 2 differs from method 1.

Note also that the original missing 'contrast', method 2 vs method 3, is now available in `fit3`. We see that method 3 is markedly worse than method 2.

Student language score by teaching method and IQ. . .

Interpreting the output. . .

Let us put confidence bounds on our effects.

```
> ## Baseline method here is method1.
> confint(TeachIQmethod.fit2)
                 2.5 %     97.5 %
(Intercept)  24.1218063 60.0492251
IQ            0.1647361  0.4665482
method2       4.0799363 15.6759248
method3     -20.0224212 -8.2960209
```

```
> ## Baseline method here is method2.
> confint(TeachIQmethod.fit3)
                 2.5 %      97.5 %
(Intercept)  35.0127936  68.9140989
IQ            0.1647361   0.4665482
method1     -15.6759248  -4.0799363
method3     -29.7496781 -18.3246250
```

Student language score by teaching method and IQ. . .

An Executive Summary

The best teaching method is method 2 regardless of the student's IQ.

With 95% confidence, we can make the following statements:

1. For students experiencing the same teaching method, we estimate that the mean language test score increases by between 1.6 and 4.7 marks for each additional ten[7] IQ points.
2. For students with the same IQ, we estimate that the mean language test score. . .
   - for students taught using method 2 is between 4.1 and 15.7 marks higher than those taught using method 1,
   - for students taught using method 3 is between 8.3 and 20 marks lower than those taught using method 1,
   - for students taught using method 3 is between 18.3 and 29.7 marks lower than those taught using method 2.

[7] This is the CI for $\beta_1 \times 10$ to demonstrate stating it on a different scale.

Section 9.4
Closing remarks and relevant R-code.

Most of the R-code you need for this chapter

When your response variable can be explained by a categorical variable and a numeric variable then you can use both explanatory variables in your analysis to explain y.

You do not need to create indicator variables for levels of the categorical variable since R does this for you. It will choose the reference level for you, so be careful. You can change this if needed. For example:

```
> teach.df$method = relevel(teach.df$method, ref = "2")
```

Most of the R-code you need for this chapter...

Follow the following steps:

Gain some intuition from looking at a suitable scatter plot of the data – this could suggest whether the response should be logged, etc.

Then, fit the model with interaction term,

```
> TeachIQmethod.fit=lm(lang~IQ*method, data=teach.df)
```

Then, assuming no reason to question the independence assumption, do our usual EOV, normality and influence checks.

If checks are OK, then see whether you really need a separate slope for each level of your factor variable. We can use `summary` if the factor variable has two levels, otherwise we must use `anova`

```
> data.frame(anova(TeachIQmethod.fit))
```

Most of the R-code you need for this chapter...

If the interaction term is not significant (large *P*-value for the ‘:’ term) then apply Occam’s razor by fitter the simpler main effects model.

To fit the simpler ‘main effects’ (parallel lines) model:

```
> TeachIQmethod.fit2=lm(lang~IQ+method, data=teach.df)
> summary(TeachIQmethod.fit2)
```

Note that sometimes we may see that we don’t need one or more of the explanatory variables at all – do this by checking for a large *P*-value on each term. Apply Occam’s razor accordingly.

Cautionary remark about anova

When using anova to test a joint hypothesis for a multi-level explanatory factor, that factor **must be the bottom term** in the anova output.[8] That is, immediatately above the Residuals line.

By way of example, suppose we change the order of the explanatory variables when we fit the model using lm,

```
> TeachIQmethod.fit2b=lm(lang~method+IQ, data=teach.df)
> anova(TeachIQmethod.fit2b)
```

```
          Df Sum Sq Mean Sq F value    Pr(>F)    
method     2 3194.6 1597.30  41.490 8.112e-09 ***
IQ         1  711.6  711.65  18.485 0.0002134 ***
Residuals 26 1001.0   38.50                      
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

In this case, the change in *P*-values (see slide 23) makes no difference to conclusions , but this may not always be the case.

[8] Unlike the summary table, the *P*-values for the terms in the anova output depend on the terms above it.