ANOVA

Download the `EssayMarks` dataset from my webpage:

```
> file_path <- "http://www.statslab.cam.ac.uk/~rds37/teaching/statistical_modelling/"
> (EssayMarks <- read.csv(paste0(file_path, "EssayMarks.csv")))
```

The data was collected as follows: 60 male (psychology) undergraduates read an essay supposedly written by a female undergraduate. They then evaluated the quality of the essay. By means of a photo attached to the essay, 20 students were led to believe that the writer was physically attractive and 20 that she was unattractive. The remaining students read the essay without any information about the writer's appearance (this was the control group). Within each of these three groups of 20 students, half read a version of the essay that was well written while the other half read a version that was poorly written. The aim of the experiment was to see whether the attractiveness of female students affected the male students' perception of the quality of their work.
Notice that in the dataframe `EssayMarks` that you have created, the columns `Quality` and `Attractiveness` are categories. These variables have a special type known as a factor in R. Conceptually, factors are variables in R which take on a limited number of different values called *levels*. To get a better understanding of them, let us create a vector of factors with the `factor` function.

```
> fdata <- factor(c(1, 2, 3, 4, 3, 2))
> fdata
[1] 1 2 3 4 3 2
Levels: 1 2 3 4
```

We can both see and modify the levels of a factor with the `levels` function:

```
> levels(fdata)
[1] "1" "2" "3" "4"
> levels(fdata) <- c("Fisher", "Bayes", "Neyman", "Pearson")
> fdata
[1] Fisher  Bayes   Neyman  Pearson Neyman  Bayes
Levels: Fisher Bayes Neyman Pearson
```

Turning to our data, we view the levels of the factors for the first two variables.

```
> attach(EssayMarks)
> levels(Quality)
[1] "Good" "Poor"
> levels(Photo)
[1] "Attractive"   "Control"      "Unattractive"
> Photo <- relevel(Photo, "Control")
> levels(Photo)
[1] "Control"      "Attractive"   "Unattractive"
```

The penultimate command simply re-orders the levels of `Photo` so they are more suitable for analysis. R will use corner point constraints for the parameters by default, and will enforce that the coefficient for the first level in each factor is set to zero.
Now let us do some Statistics! First let us plot the data to get a feel for it

```
> plot(Mark ~ Quality)
> plot(Mark ~ Photo)
```

As expected, the poor quality essay has been awarded lower marks on average (though the central line in the box plots gives the median, not the mean of the data). Unfortunately, it also appears that the essays attached with the "Unattractive" photo were awarded lower marks. Can this be explained by random variation or is this a real relationship that we are seeing?

Let Y_{ijk} be the (random variable representing) the mark awarded by the k^{th} student ($k = 1, \ldots, 10$) in the group that was given an essay of the i^{th} quality type ($i = 1, 2$ corresponding to "Good" and "Poor" quality respectively) and j^{th} photo type ($j = 1, 2, 3$ corresponding to "Control", "Attractive" and "Unattractive" respectively). We consider three models:

$$Y_{ijk} = \mu + \alpha_i + \varepsilon_{ijk} \tag{1}$$

$$Y_{ijk} = \mu + \alpha_i + \beta_j + \varepsilon_{ijk}, \tag{2}$$

$$Y_{ijk} = \mu + \alpha_i + \beta_j + \gamma_{ij} + \varepsilon_{ijk}, \tag{3}$$

where in each case, the ε_{ijk} are i.i.d. $N(0, \sigma^2)$. To ensure that the models are identifiable (i.e. that different combinations of parameter values give rise to different distributions for the Y_{ijk}), we will impose the corner point constraints $\alpha_1 = \beta_1 = \gamma_{1j} = \gamma_{i1} = 0$.
What does the intercept term μ represent in each of the models above? In model (1), it gives the mean mark awarded when the essay is "Good". In models (2) and (3), it is the mean mark awarded when the essay is "Good" and no photo is attached, so we are in the "Control" group. The coefficients β_2 and β_3 in model (2) give the differences in mean mark when the photo is "Attractive" and "Unattractive" respectively, compared to the "Control" group (for any fixed value of "Quality").
Model (2) is a balanced additive two-way ANOVA. In contrast, model (3) includes an interaction term, γ_{ij}. Note that this model allows for any combination of mean marks for the six experimental conditions (arising from each possible pair of "Quality" and "Photo" levels).
We can fit each of these models with the following code.

```
> EssayMarksLM1 <- lm(Mark ~ Quality)
> EssayMarksLM2 <- lm(Mark ~ Quality + Photo)
> EssayMarksLM3 <- lm(Mark ~ Quality*Photo)
```

Try `?formula` to understand more about the syntax. Note the final model includes an interaction term. It can equivalently be written as `lm(Mark ~ Quality + Photo + Quality:Photo)`.
How has R dealt with the factor variables? To understand this try applying `model.matrix` to each of the `lm` outputs e.g. `model.matrix(EssayMarksLM1)`. This gives the design matrices used in the fits. Notice how R has used our corner point constraints by default. Examine the output of the summary function applied to each of the models fitted e.g.

```
> summary(EssayMarksLM2)

Coefficients:
                  Estimate Std. Error t value Pr(>|t|)
(Intercept)        18.0332     1.1467  15.726  < 2e-16 ***
QualityPoor        -4.7663     1.1467  -4.157 0.000112 ***
PhotoAttractive     0.7495     1.4044   0.534 0.595673
PhotoUnattractive  -3.5500     1.4044  -2.528 0.014327 *
---
Signif. codes:  0 *** 0.001 ** 0.01 * 0.05 . 0.1   1

Residual standard error: 4.441 on 56 degrees of freedom
Multiple R-squared:  0.3331,Adjusted R-squared:  0.2974
F-statistic: 9.325 on 3 and 56 DF,  p-value: 4.269e-05
```

Note we have abbreviated the output. Why are there no `QualityGood` or `PhotoControl` coefficients?
The output here suggests that the mean mark when the photo is attractive is not significantly different to the mean mark for the control group (though the coefficient is positive). On the other hand, the mean mark when the photo is unattractive is significantly different to the control group, and lower. Rather depressingly, this effect appears to have a name in the psychology literature: you can try Googling the "halo effect".
We can formally test whether the parameters for the type of photo are collectively significant using an F-test implemented by

```
> anova(EssayMarksLM1, EssayMarksLM2)
```

To graphically gauge whether a model with interactions may be appropriate, we can use the `interaction.plot` function:

```
> interaction.plot(Photo, Quality, Mark)
> interaction.plot(Quality, Photo, Mark)
```

If interactions were not needed, we would expect the lines of these plots to be roughly parallel (why?). Indeed, the fitted values of the additive two-way ANOVA yield parallel lines here:

```
> interaction.plot(Photo, Quality, fitted.values(EssayMarksLM2))
> interaction.plot(Quality, Photo, fitted.values(EssayMarksLM2))
```

To test for the presence of interactions, we again use an F-test:

```
> anova(EssayMarksLM2, EssayMarksLM3)
```

Make sure you understand the output here—it may help to refer to practical 4. Which of the three models seems to be the most appropriate for the data?
From the interaction plots, it appears that the presence of the attractive photo does not affects the mean marks much, but the mean marks with presence of the unattractive photo are rather different from the control group. It thus makes sense to try to combine the control and attractive photo groups into one:

```
> Photo_grp <- Photo
> levels(Photo_grp)
[1] "Control"      "Attractive"   "Unattractive"
> levels(Photo_grp) <- c("Control+Attr", "Control+Attr", "Unattractive")
> levels(Photo_grp)
[1] "Control+Attr" "Unattractive"
```

Exercise: Write out the models being fitted by the following code.

```
> EssayMarksLM4 <- lm(Mark ~ Quality*Photo_grp)
> EssayMarksLM5 <- lm(Mark ~ Quality + Photo + Quality:Photo_grp)
```

To compare all the models we have fitted, we can use Akaike's Information Criterion.

```
> AIC(EssayMarksLM1, EssayMarksLM2, EssayMarksLM3, EssayMarksLM4, EssayMarksLM5)
```

What appears to be the most appropriate model according to AIC?

ANCOVA

This section and the exercises that follow are optional. Download the `Cycling` data from the course webpage

```
> detach(EssayMarks)
> Cycling <- read.csv(paste0(file_path, "Cycling.csv"))
> str(Cycling)  # You can see which variables are factors and how many levels they have
> attach(Cycling)
```

These data were collected by Prof. Ian Walker from the University of Bath. He used an instrumented bicycle to gather proximity data from overtaking motorists when cycling. Recorded in the data is the distance from kerb when a car passed, the type of road that he was cycling on, which city he was in, whether or not a helmet was being worn and other variables. The goal of this data collection was to determine whether wearing a cycle helmet affects how close motorists pass by cyclists.

Exercises

1. Fit a linear model to the data with `passing.distance` as the response. Notice that the Q–Q plot looks rather suspect. You could try using the `boxcox` function in the MASS package to suggest a transformation of the data (perhaps try using a cube root transformation, though this doesn't improve things much). However, as the sample size is reasonably large, we might expect that even if normality of the errors does not hold, our conclusions should not be too erroneous.

2. What appears to be the effect of wearing a helmet? Note that though we have a *statistically* significant result, the effect of wearing a helmet is rather small and most probably *insignificant* from a practical point of view, particularly when compared to the fact that the helmet could save your life if you are unfortunate enough to have an accident.

3. Try to find a smaller model that adequately explains the data, either using `stepAIC` form the MASS package, or manually performing different F-tests using the `anova` function.