MATH3012 – Statistical Methods II
S-Plus Worksheet 4 – Analysis of Covariance

The dataset `birth` contains data on the weight of 24 newborn babies. There are two explanatory variables; sex (S; a qualitative variable, coded "1=male" and "2=female") and gestational age (X; a quantative variable, in weeks) together with the response variable, birthweight (Y; in grams).

These data can be plotted, with males and females distinguished, using

```
> plot(birth$x,birth$y, xlab="Age", ylab="Birthweight", las=1, type="n")
> points(birth$x[birth$s==1],birth$y[birth$s==1])
> points(birth$x[birth$s==2],birth$y[birth$s==2],pch=0)
```

S is a factor (the labels 1 and 2 have no intrinsic meaning), and will need to be be declared as such using the `factor` command. Alternatively, you can recode the levels of S to `M` and `F`, which has the same effect.

We can now fit models which include both qualitative and quantitative explanatory variables. Such models are sometimes called *Analysis of Covariance* models. For example

```
> birth.lm1 <- lm(y~s+x,data=birth)
```

fits the model

$$Y_i = \mu + \alpha(s_i) + \beta x_i + \epsilon_i \qquad i = 1, \ldots, n \tag{1}$$

where s_i is the level of S and x_i the corresponding observation of X for the ith observation. Provided that we have set `options(contrasts=c("contr.treatment", "contr.poly"))`, $\alpha(s)$ is constrained to be equal to 0 at the first level of S. The interpretation of this model is that $E(Y)$ depends linearly on X, and that the linear relationship has a different intercept for different levels of S. The parameter $\alpha(s)$ describes the differences in the intercepts.

Try removing terms from this model. If we remove S, the model is just a linear regression of Y on X, and allows no dependence of Y on S. If we remove S, then $E(Y)$ is the same for any value of X, and just differs between the levels of S.

Model (1) states that S and X afect Y but that they do so independently. The difference between $E(Y)$ at two different values of X is the same for every level of S. Similarly, the difference between $E(Y)$ at two different levels of S is the same for every value of X. However, we can also incorporate interaction between a factor and a quantitative explanatory variable.

```
> birth.lm2<- lm(y~s+x+s:x,data=birth)
```

fits the model

$$Y_i = \mu + \alpha(s_i) + \beta x_i + \gamma(s_i) x_i + \epsilon_i \qquad i = 1, \ldots, n \tag{2}$$

where $\alpha(s)$ and $\gamma(s)$ are constrained to be equal to 0 at the first level of S. The interpretation of this model is that $E(Y)$ depends linearly on X, and that the linear relationship has a different intercept *and a different slope* for different levels of S. The parameter $\gamma(s)$ describes the differences in the slopes.

Again, we can compare nested models using F-tests, and again it only makes sense for an interaction to be present if all the corresponding main effects are also present.

Which model do you feel best describes the relationship between birthweight, sex and gestational age? For your chosen model plot the data as above, and superimpose the regression line (or lines if different for males and females) on your plot; see Worksheet 1 for details. Hence interpret your model. Is it a good fit?