MSc Business Analytics: **Quantitative Methods**

Exercises - SOLUTIONS

1.

(a) The scatterplot matrix indicates that the strongest linear relationships with the response variable individually are with `weight` and `mat.weight` (or `mat.height`). These relationships are positive, so that `dl.milk` increases with increasing values of these variables. There would appear to be an effect for `sex` (breast milk intake being less for girls), but this is not clear from the plot. There is little relationship, if any, between the response and `ml.suppl`.

These comments are reinforced by the individual regressions, which show significant positive relationships with `weight`, `mat.weight` and `mat.height`, at the 1% level. `sex` is also significant, at the 5% level.

Candidates for inclusion in a good multiple linear regression model are `weight` and `mat.weight` and `mat.height` (and possibly `sex`). Unsurprisingly, the largest correlation between the explanatory variables is between `mat.weight` and `mat.height`, so perhaps there would be no need for both these terms in the model (although the correlation 0.565 is not all that large).

(b) (i) Since the model includes an intercept it would be over-parameterized if coefficients for both `sexvar==1` and `sexvar==0` were included. For this reason R creates a dummy variable for `sexvar` which assigns the value 0 to `boy` and 1 to `girl`. Hence the coefficient of `sexvar` gives the change in intercept for a `girl` over the model baseline, which is given in reference to a `boy`.

[Note: this is confirmed by the `dummy.coef` output].

(ii) The model gives the expected breast milk intake in 24 hours as:

$$\widehat{\texttt{dl.milk}} = -11.6818 - 0.4995\ \texttt{sexvar} + 1.3491\ \texttt{weight} - 0.0022\ \texttt{ml.suppl} + 0.0062\ \texttt{mat.weight} + 0.0723\ \texttt{mat.height}$$

Hence, for the given values we find (`sexvar=0`)

$$\hat{y} = -11.6818 + 1.3491(5.5) - 0.0022(0) + 0.0062(60) + 0.0723(168) = 8.25665$$

or 8.26 dl/24hr.

(iii) The cook distance is given as

$$D_i = \frac{e_i'^2}{(p+1)} \frac{h_i}{(1-h_i)}$$

Hence, point 32 has a large cook statistic as it has both high leverage (≈ 0.45, indicating this point has the *potential* to influence) and a high standardized residual (indicating a poor fit to the model, ≈ 2). Note that the leverage is $> 2p/n = 12/50 = 0.24$, the *rule of thumb* value. Only one other point (10) has a value above this figure, but its standardized residual is much less.

(c) (i) In Model B `weight`, `mat.height`, `ml.suppl` and `sexvar` are chosen as explanatory variables. This is not that surprising given the earlier comments. We may have expected `mat.weight` to be a better predictor - perhaps this is not a reliable measure post pregnancy?).

(ii) Model A was seen to have redundant terms, so an equally well fitting model with less term(s) is certainly possible.

(iii) The assumptions underlying the (multiple) linear regression are that the error terms are (independently) normally distributed with zero mean and constant variance. These assumptions can be checked by reference to the plots of standardized residuals from the model.

The probability (qq-) plot shows no reason for concern regarding the normality assumption. Also, the plot of standardized residuals shows no apparent pattern - the residuals being evenly scattered around zero. However point 32 has still a large cook statistic as it has both high leverage and a high standardized residual.

(d) We could fit a series of regressions using `dl.milk` as the response and `ml.suppl` and the factor (/dummy variable) `sex` as explanatories.

There are three models to compare:

mod1: $\texttt{dl.milk} \sim \texttt{ml.suppl}$
mod2: $\texttt{dl.milk} \sim \texttt{ml.suppl} + \texttt{sex}$
mod3: $\texttt{dl.milk} \sim \texttt{ml.suppl} + \texttt{sex} + \texttt{sex} : \texttt{ml.suppl}$

mod 1 fits a single regression line between `dl.milk` and `ml.suppl` so that the relationship is assumed the same for each sex; mod2 allows each sex to have their own intercept, and mod3 allows for separate intercepts and slopes.

Since the models are nested, the terms of the fitted models can be compared to see whether the separate models (mod2 or mod3) add anything over the simple linear regression of mod1.

[This could be tested formally using, for example,

```
> anova(mod1 ,mod2, mod3)
```

or by working backwards from mod3, checking for the significance of the additional terms].