MATH 523 - EXERCISES 1

These exercises are not for assessment

1 The `birthwt` data set included in the `MASS` library in the package `R` contains information on the low birthweight status of 189 babies born at Baystate Medical Center, Springfield, Mass during 1986. The recorded variables are

- `low`: indicator of birth weight less than 2.5 kg.
- `age`: mother's age in years.
- `lwt`: mother's weight in pounds at last menstrual period.
- `race`: mother's race (1 = white, 2 = black, 3 = other).
- `smoke`: smoking status during pregnancy.
- `ptl`: number of previous premature labours.
- `ht`: history of hypertension.
- `ui`: presence of uterine irritability.
- `ftv`: number of physician visits during the first trimester.
- `bwt`: birth weight in grams.

It is possible to perform a **linear regression analysis** for these data, taking the **binary** variable `low` as the response variable. Using this response variable, and standard techniques from linear regression modelling, explore which of the predictors are useful in predicting low birthweight, and comment on the general suitability of your selected model for performing the prediction.

Repeat the exercise using the continuous variable `bwt` as the response, and compare results.

2 The data set contained in the file **Absences.csv** (available on the course website) records the numbers of days that high-school juniors are absent during a given time period at two schools. The data set comprises 316 children: the recorded variables are

- `id`: child identification number.
- `school`: factor indicating school.
- `male`: factor indicating sex of child (1=male).
- `math`: standardized test scores for mathematics.
- `langarts`: standardized test scores for language arts.
- `daysatt`: number of days attended.
- `daysabs`: number of days absent.

The response variable is `daysabs`. Using linear regression techniques, explore models that might be used to predict this response, and assess the quality of fit. Identify any potential drawbacks of using the linear regression approach for these data.

3 Suppose that $\mathbf{X}_1$ $(n \times p_1)$ and $\mathbf{X}_2$ $(n \times p_2)$ and are two portions of a design matrix $\mathbf{X} = [\mathbf{X}_1\ \mathbf{X}_2]$ $(n \times p)$ where $p = p_1 + p_2$. Suppose that the model

$$\boldsymbol{Y} = \mathbf{X}\beta + \epsilon$$

with $\text{Var}[\epsilon] = \sigma^2\mathbf{I}_n$ is to be fitted, but that in reality, the variables in $\mathbf{X}_2$ have **no** influence on response $\boldsymbol{Y}$; that is, the true value of β is the $(p \times 1)$ vector $(\beta_{10}, \mathbf{0}_{p_2})^\mathsf{T}$, where $\mathbf{0}_q$ denotes a vector of q zeros.

Establish the properties of the least-squares estimator of β derived using the **fitted** model. Compare this with properties of the corresponding estimator if the $\mathbf{X}_2$ variables are omitted from the model. In particular, comment on the variance of the estimator of β_1, the coefficients corresponding to the $\mathbf{X}_1$ variables.