STAT 210
Applied Statistics and Data Analysis
Problem List 10
(Due on Week 11)

We will use the data in the file `iguanodon.csv` for the first two problems in the list. The data come from measurements taken in Jurassic Park and has information on 27 specimens of a type of dinosaur known as iguanodon. Read the data onto a data frame. There are nine variables in the set:

- `length`, the length from head to tail in m,
- `weight`, the weight in tons,
- `width`, the width in m,
- `height`, the height in m,
- `leg_length`, the average length for the two (rear) legs in m,
- `arm_length`, the average length for the two arms in m,
- `head_length`, the length of the head in cm,
- `species`, the species with two values, `A` and `B`,
- `power`, the strength index for the dinosaur, and
- `bite_st`, the bite strength in normalized units.

Problem 1

In this question you have to explore the relationship between the variables `bite_st` and `head_length`.

(a) Graph a scatterplot of `bite_st` as a function of `head_length`. Fit a simple regression model for these variables and add the regression line to the plot. Comment on the plot. Print the summary table. What is the R^2 for this model? Write down the equation for the regression line, including all the terms, and give an interpretation of the parameters. Predict the bite strength of an iguanodon with a head length of 90 cm. and include a prediction interval at the 98% confidence level.

(b) State clearly the assumptions on which the linear regression model is based. Use graphical methods and tests to check these assumptions. What are your conclusions?

(c) There are two species of iguanodons in the file, denoted `A` and `B`, and this characteristic is available in the categorical variable `species`. If this variable was not read as a `factor`, transform it before you continue. Fit a model that includes `head_length`, `species`, and the interaction between the two. Using a critical value for α of 0.05 and starting with the complete model, select a minimal adequate model. Compare the adjusted R^2 with the previous model. Check the assumptions for the final model.

(d) Write down the equation for the regression model in (c) and predict the value of the bite strength for iguanodons of both species having head length 90 cm, including prediction intervals at the 98% confidence level. Compare with the previous prediction and comment.

Problem 2

In this question you have to develop a model for `power` as a function of the numerical variables in the set, excluding `bite_st`.

(a) Do a scatterplot matrix for the numerical variables in the data set, excluding `bite_st`. Calculate and graph the correlation matrix for these variables. Comment on the results.

(b) Fit a regression model for `power` as a function of the variables mentioned in (a). With a threshold for the variance inflation factor of 2, use a sequential procedure to eliminate variables that may cause multicollinearity problems.

(c) Using a backward selection procedure with a critical α of 0.10 and starting with the variables you selected in (b), obtain a minimal adequate model. Comment on the steps that you take.

(d) Fit a model using the BIC criterion starting with the variables you selected in (b). Compare your final model with the result of (c).

(e) Write an equation for the final model in (c) and interpret the coefficients. Predict the `power` for an iguanodon with the following covariates. Include confidence intervals at the 99% level.

Table 1: Covariates for prediction

`length`	`weight`	`width`	`height`	`leg_length`	`arm_length`	`head_length`
10	5.1	5.2	7.3	1.4	1.3	89

(f) Print an anova table for the final model and find the estimated standard deviation of the errors. Describe explicitly the sampling distribution for the estimated parameters.

Problem 3

Using the `sat` dataset in the `faraway` package, fit a model with the `total` SAT score as the response and `expend`, `salary`, `ratio` and `takers` as predictors. Perform regression diagnostics on this model to answer the following questions. Display any plots that are relevant. Do not provide any plots about which you have nothing to say.

(a) Check the constant variance assumption for the errors.
(b) Check the normality assumption.
(c) Check for large leverage points.
(d) Check for outliers.
(e) Check for influential points.

Problem 4

For this question, use the data set `uscrime` in the package `HH`. After loading the library, you need to run `data("uscrime")`. Do not mistake with `UScrime`. For this exercise, values for the variance inflation factor (vif) below 5 are considered acceptable. The following commands load the data:

```
library(HH)
data("uscrime")
```

(a) Fit a multiple regression model for `R` using all the other variables except `State`. Look at the summary and variance inflation factors and **comment**.

(b) Use the function `stepAIC` in package `MASS` to get a reduced model. Get information about this model using `summary` and look at the variance inflation factors. **Comment on these results**.

(c) Starting with the model produced in (b), drop, one by one, any variables that have a vif greater than 5 or non-significant p-value (use $\alpha = 0.05$). Give a summary of your final model and write down the corresponding equation.

(d) Check the validity of the model assumptions starting with diagnostics plots, and carry out any necessary tests. **Comment all your steps.**