FALL 2010
Stat 471: Homework Assignment 2
Due: October 8, 2010

1. Install the `EngrExpt` package that contains the data sets from the textbook *Introductory Statistics for Engineering Experimentation.* Read the documentation for the `timetemp` data in which the response, `time`, is measured along with one continuous covariate, `temp`, and one categorical covariate, `type`.

   (a) Plot the data in one panel of `temp` versus `time` with different colors for the Repair and for the OEM typess. The documentation contains code for producing such a plot using the lattice package. Try to create a similar plot with ggplot2. Add the two simple linear regression lines for the two types. (In ggplot2 you can append `+ geom_smooth(method="lm")` to the call to `qplot`.)

   (b) Plot the data in two panels of `temp` versus `time`: one panel for the Repair type and one panel for the OEM type. Add the regression lines to each panel of the plot. (Same code as above).

   (c) Our purpose is to decide:

      i. Does each subgroup of the data follow a straight line pattern (more or less)?

      ii. Are the two lines coincident? (That is, could I fit all the data with one line, ignoring the `type` variable?)

      iii. Are the two lines parallel? (That is, I allow for different intercepts but require the same slope.)

      iv. Is there sufficient evidence to require two lines with different slopes and different intercepts?

      Answer these questions based on the plots you have created (i.e. before fitting models). Which plot is more useful in examining these questions and why?

   (d) Fit the following three models to these data using the `lm()` function:

   ```
   time ~ 1 + temp
   time ~ 1 + type + temp
   time ~ 1 + type + temp + time:temp
   ```

   For each of the three fitted models, explain which of the situations described above (coincident lines, distinct parallel lines, distinct non-parallel lines) is being modeled.

   (e) For each model state the numerical value of the estimated intercept and of the estimated slope for the Repair and for the OEM type.

   (f) Suppose you called the fitted models `lm1`, `lm2` and `lm3`. What does a comparative analysis of variance produced by `anova(lm1,lm2)` test about the lines? There is a t-test in one of the coefficient tables that corresponds to this F test. Which one is it? Answer these two questions (what is being tested and which t-test is equivalent) for the results of `anova(lm2,lm3)`.

   (g) Suppose we adopt the traditional criterion that a p-value less then 5% is regarded as "significant". Would you prefer model `lm3` to model `lm2`? What about `lm2` compared to `lm1`.

2. Fit, using `lm()`, the model

   ```
   log(Volume) ~ 1 + log(Girth) + log(Height)
   ```

   to the `trees` data. (It is part of the `datasets` package, which will already be installed for you.)

(a) Simulate 10000 response vectors from this model (remember to first call `set.seed()` with some integer value so I can reproduce your results) with the fitted values of the parameters (use the `simulate()` function to produce a matrix of simulated response vectors). Then fit these responses to the model shown above (use the `lm()` function with the name of the matrix of responses on the left hand side of the formula).

(b) Extract the coefficients, which should be in the form of a matrix with 3 rows and 10,000 columns. Create empirical density plots for each of the coefficients (i.e. each row of this matrix). Do these values appear to be normally distributed? Does the mean and standard deviation of each row agree with the estimate and standard error of that parameter in the original model fit?

(c) For each row determine the central interval in which 95% of the estimates fall. (Recall that you use an expression like

```
> int1 <- quantile(myCoefs[1, ], c(0.025, 0.975))
```

to do this for each row.) Compare these to the confidence intervals on the coefficients for the original model fit. If everything has gone right they should be close. Are they?