Statistics 203
Introduction to Regression and Analysis of Variance
Assignment #2
Due Thursday, February 10

Prof. J. Taylor

USE R FOR ALL CALCULATIONS. PROVIDE COPIES OF YOUR CODE IN THE ASSIGNMENT.

Q. 1) The dataset http://www-stat/~jtaylo/courses/stats203/data/education.table contains data relating to the per-capita education expenditure `education` by state in 1975. The variables include: `income`, per capita income in 1973; `under18`, the proportion of population under 18; `urban`, the proportion living in urban areas and `region`, a qualitative variable describing a geographic grouping of the states.

(a) Using stepwise or all subsets regression, select an appropriate multiple linear regression model to the above data, using whatever interactions you feel appropriate.

(b) After finding an appropriate model, verify, with appropriate plots that the regression model is adequate using diagnostic plots and/or added variable plots.

Q. 2) In the treatment of HIV, it is of interest to determine whether the genotype of a patient's HIV virus can be used to decide on what treatment a patient should receive if he/she is failing their current therapy. Towards this goal, a scoring system known as the "Genotypic Sensitivity Score" (GSS) (for more information see http://hivdb.stanford.edu). The data found at

http://www-stat/~jtaylo/courses/stats203/data/vl.table

consists of two columns: one the GSS of a population of patients, the other their viral load (a measure of how much virus is in the blood) at a future time point.

(a) Plot the data: are there any outlying observations? Based on the plot of the data, is it reasonable to test for the presence of outliers using `outlier.test`?

(b) For this question, we will remove the outliers in the data. Fit a simple linear regression model to the data (after removing the outlier) with `VL` as the outcome and `GSS` as the predictor.

(c) Does the data (with outliers removed) satisfy the usual regression assumptions? Report any diagnostics plots you used to answer this question, as well as any violations to the usual assumptions.

Q. 3) (*ALSM*, 19.14)

A research laboratory was developing a new compound for the relief of severe cases of hay fever. In an experiment with 36 volunteers, the amounts of the two active ingredients (factors A and B) in the compound were varied at three levels each. Randomization was used in assigning four volunteers to each of the nine treatments. The data can be found at:

http://www-stat/~jtaylo/courses/stats203/data/hayfever.table.

(a) Fit the two-way ANOVA model, including interactions. What is the estimated mean when Factor A is 3 and Factor B is 2?

(b) Using `R`'s standard regression plots, plot the QQplot of the residuals. Is there any serious violation of normality?

(c) This question asks you to graphically summarize the data. Create a plot with Factor A on the x-axis, and, using 3 different plotting symbols, the mean for each level of Factor B above each level of Factor A. Does there appear to be any interactions?

(d) Test for an interaction at level $\alpha = 0.05$.

(e) Test for main effects of Factors A and B.

Q. 4) (*Two-Way ANOVA, equal sample sizes*)

(a) Work out the expected mean squares of each term in the two-way ANOVA table below, from the model

$$Y_{ijk} = \mu + \alpha_i + \beta_j + (\alpha\beta)_{ij} + \varepsilon_{ijk}, \quad 1 \leq i \leq r, 1 \leq j \leq m, 1 \leq k \leq n.$$

(b) Show that

$$SSR = SSA + SSB + SSAB.$$

Term	SS
A	$SSA = nm\sum_{i=1}^{r}\left(\overline{Y}_{i\cdot\cdot} - \overline{Y}_{\cdots}\right)^2$
B	$SSB = nr\sum_{j=1}^{m}\left(\overline{Y}_{\cdot j\cdot} - \overline{Y}_{\cdots}\right)^2$
AB	$SSAB = n\sum_{i=1}^{r}\sum_{j=1}^{m}\left(\overline{Y}_{ij\cdot} - \overline{Y}_{i\cdot\cdot} - \overline{Y}_{\cdot j\cdot} + \overline{Y}_{\cdots}\right)^2$
Error	$SSE = \sum_{i=1}^{r}\sum_{j=1}^{m}\sum_{k=1}^{n}(Y_{ijk} - \overline{Y}_{ij\cdot})^2$