## Homework 6

1. Let $\epsilon_{ij}$, $i = 1, 2, \ldots, I$, $j = 1, 2, \ldots, J$ be independent $N(0, \sigma^2)$. Prove that $\sum_i \sum_j (\bar{\epsilon}_{i.} - \bar{\epsilon}_{..})^2$ and $\sum_i \sum_j (\bar{\epsilon}_{.j} - \bar{\epsilon}_{..})^2$ are statistically independent.

   By calculating that the covariance is 0, verify that $X_i - \bar{X}$ and $\bar{X}$ are independent if $X_i \sim_{iid} N(\mu, \sigma^2)$. This shows that $\bar{\epsilon_{i.}} - \bar{\epsilon_{..}}$ and $\bar{\epsilon_{..}}$ are independent. Next

$$\begin{aligned} Cov(\bar{\epsilon_{i.}} - \bar{\epsilon_{..}}, \bar{\epsilon_{.j}}) &= Cov(\bar{\epsilon_{i.}}, \bar{\epsilon_{.j}}) - Cov(\bar{\epsilon_{..}}, \bar{\epsilon_{.j}}) \\ &= \frac{1}{IJ}\sigma^2 - \frac{1}{IJ^2} J\sigma^2 \\ &= 0 \end{aligned}$$

   We thus have that the vector with components $\bar{\epsilon_{i.}} - \bar{\epsilon_{..}}$ is independent of the vector with components $\bar{\epsilon_{.j}} - \bar{\epsilon_{..}}$, since zero covariance implies independence for normals. Finally, if $U$ and $V$ are independent, so are $g(U)$ and $f(V)$.

2. The "jointed line" (also called "broken stick") regression model is as follows. (See section 7.2.1 of Faraway.) Suppose that for $x \leq \xi$, $E(Y|X = x) = \alpha_1 + \beta_1 x$, that for $x > \xi$, $E(Y|X = x) = \alpha_2 + \beta_2 x$, and that the regression function is continuous at $\xi$.

   (a) Show that the set of all such functions is a linear space — any linear combination of them is such a function.

   A constant times any such function is of that type. Adding two such functions together gives another of that type, since the sum will be linear in the two regions and will also be continuous.

   (b) There are a number of basis sets for this space. Show that the following functions are a basis: $f_1(x) = 1$, $f_2(x) = x$, $f_3(x) = (x - \xi)I\{x \geq \xi\}$. Show how the function could be expressed as a linear model with respect to this basis

   It is easy to see that the functions are linearly independent. Suppose that $g(x)$ is of the form described above. Then

$$g(x) = \alpha_1 f_1(x) + \beta_1 f_2(x) + (\beta_1 - \beta_1) f_3(x)$$

3. Old Faithful geyser in Yellowstone National Park, Wyoming, derives its name from the regularity of its eruptions. The file `oldfaithful.csv` contains measurements on eight successive days of the durations of the eruptions (in minutes) and the subsequent time interval before the next eruption. The park posts predicted eruption times for vistors. How well can the time until the next eruption be predicted by the duration of the current one?

   (a) Plot time intervals versus duration of the previous eruption. Regress intervals on duration and plot the residuals. Comment briefly on the fit.

   There plot shows two clusters. Also, a plot of duration versus event number shows that the time series essentially oscillates alternately back and forth between these two levels.

   (b) Observe that there may be two regimes, corresponding to short and long eruptions, so that perhaps the regression should be segmented as a "broken stick" regression. Choose a breakpoint by eye, fit a segmented regression, and plot the result. Compare to the result of fitting a single regression.

   (c) Use an F test to compare the segmented regression to the single regression.

   One way to do the test is to use RSS for the two models, with and without the breakpoint. The test rejects.

   (d) The analysis above ignores possible temporal trends. Plot the residuals versus time and comment.

   There may be a suggestion of a week daily trend, but it is not obvious to me. The most striking time dependence is that mentioned in (a) above.

   (e) How could the breakpoint $\xi$ be estimated other than by eye?

   The function could be fit by least squares. The residual sum of squares is nonlinear in $\xi$, but for each choice of $\xi$ the other three parameters can be estimated as in part (b). A simple method would thus be to search for the minimizing $\xi$ on a fine grid.