Homework 3, Stat 3

Use Rmarkdown for answers to problems 3 and 4, documenting codes, outputs and your explanations/formulation etc.

1. Consider the following model:

$$Y = X_1\beta_1 + X_2\beta_2 + \epsilon$$

   where Y and ϵ are n dimensional vectors, X_1 and X_2 are $n \times p$ and $n \times q$ dimensional matrices of predictors, β_1 and β_2 are unknown regression coefficient vectors of dimensions p and q.

   Let SSR_1 be the sum of squares residuals from the model and SSR_2 be the sum of squares residuals when $\beta_2 = 0$. Assume that ϵ has a multivariate normal distribution with mean zero and variance $\sigma^2 I$.

   (a) Show that SSR_1 follows a χ^2 distribution. What are the degrees of freedom?

   (b) Under the hypothesis $H_0 : \beta_2 = 0$ show that SSR_2 follows a χ^2 distribution. What are the degrees of freedom?

   (c) Under H_0, show that SSR_2-SSR_1 follows a χ^2_q distribution and is independent of SSR_1.

   (d) Form an F statistic to test H_0.

2. (Quiz 2 question) Consider simple linear regression and the problem of testing $\beta_1 = 0$. We can do this in two ways.

   A Use a t -test based on $\hat{\beta}_1 \sim \mathcal{N}(\beta_1, \sigma^2/S_{XX})$

   B Use the F test in the ANOVA

   Answer the following questions in this context.

   (a) Write down an estimator for σ^2 that is independent of $\hat{\beta}_1$.

   (b) What is the distribution of the estimator in part (1)?

   (c) Write down the t-statistic mentioned in (A) and its distribution.

   (d) Show that the F-statistic in (B) is

$$\frac{S_{YY} - \sum_{i=1}^{n}(y_i - \hat{y}_i)^2}{\sum_{i=1}^{n}(y_i - \hat{y}_i)^2/(n-2)}$$

   (e) Show that the numerator in the F-statistic equals S_{XY}^2/S_{XX}.

   (f) Hence show that the F statistic in (B) is square of the t-statistic in (A).

   (g) Conclude that the two tests (A) and (B) are equivalent by expressing the t and F distributions in terms of independent normals.

3. For the earthquake data, consider depth 12, latitude 28.7 and longitude 77.1. Answer the following questions assuming a linear model with independent normal errors.

   (a) Find a 95% confidence interval for the average magnitude of earthquake at this location.

   (b) Find a 95% prediction interval for the magnitude of an individual earthquake at this location.

4. Consider the seatpos data. Carry out a model selection exercise. Present all details in intermediate steps, the final model with proper interpretations of output.