Homework 7: Due Wednesday, March 24

Before you begin, *work through* Task 4.4.1 in the text (pages 246–250). There is no extra credit for this part, but it's designed to help you become familiar with the techniques before the very similar exercises that follow.

For both Task 4.4.1 and Question 1 below, you are expected to *make the calculations yourself*, using the matrix algebra tools within R. Do not rely on the SAS or Minitab examples provided with the text, but you are allowed and encouraged to check your answers in R using `lm`.

1. Work through the problems on page 251 of the text, specifically 4.4.1, 4.4.2, 4.4.3, 4.4.6, 4.4.7, 4.4.10, 4.4.12. Note that if the model is written $\mu_Y(x1, x_2) = \beta_0 + \beta_1 x_1 + \beta_2 x_2$ and the residual standard error is σ, the population values (given elsewhere in the text) are $\beta_0 = -5$, $\beta_1 = 0.2$, $\beta_2 = -1$, $\sigma = 1.7076$. (This is really all one question — 10 points for the whole question.)

   Note added about 4.4.12. This looks like a prediction interval question, but I don't think it's intended that way, because at this stage of the text, we haven't done prediction intervals for multiple regression. Assume your estimates $\widehat{\boldsymbol{\beta}}$ and $\hat{\sigma}$ are the true population values, and then use the normal distribution to calculate the required probability.

2. This question is based on a dataset called SENIC, which you can load into R and then edit with the following commands:

```
senic=read.table("http://www.utstat.toronto.edu/~brunner/data/legal/openSENIC.data.txt")
senic=senic[,-2]
senic=na.omit(senic)
```

(2-a) Consider the model in which `infpercent` is the response ($\mathbf{y}$) variable and `nbeds, nurses, lngstay, age` are the four covariates. Show how to construct the X matrix for this dataset and explicitly calculate the following: X^TX, $X^T\mathbf{y}$, $\widehat{\beta}$, $\widehat{\sigma}$. Calculate the standard errors of the five regression parameters (including the intercept). Then, verify your results by using the `lm` and `summary` commands in R. (Except for the last part, this is also intended to be direct calculation using the matrix operations in R — be sure to state the intermediate calculations as well as the final result.)

(2-b) Compute the standardized residuals (use `rstandard`) and plot them against (i) each of the four covariates, (ii) the fitted values, (iii) the "region" variable (a command of the form `boxplot(residuals~regions)` will do this elegantly). Also show a QQ (rankit) plot of the standardized residuals. Based on these plots, do you think the model is a good fit to the data?

(2-c) A new facility is opened with the variables nbeds=531, nurses=442, lngstay=9.1, age=55. Compute (i) a 99% confidence interval, (ii) a 90% prediction interval, for the value of infpercent at this facility. Which of the two intervals is more relevant for this problem?

Hint. A command like `s1=data.frame(nbeds=531,nurses=442,lngstay=9.1,age=55)` will create a dataframe with the required information, which you can then use with the `predict.lm` function.