EXERCISE ON MULTIVARIATE NORMAL, REGRESSION

(1) Suppose

$$\begin{bmatrix} X \\ Y \end{bmatrix} \sim N\left(\begin{bmatrix} 2 \\ 1 \end{bmatrix}, \begin{bmatrix} 1 & 0 \\ 0 & 2 \end{bmatrix}\right),$$

or in words, X and Y are jointly normal, mutually independent, with X having mean 2 and variance 1 and Y having mean 1 and variance 2.

(a) What is $P[X \in (3,4), Y \in (0,1)]$? Or in words, what is the probability that X is between three and four and Y is between 0 and 1? [Hint: You should be able to answer this using only a table of the univariate normal probabilities or a computer command to evaluate normal probabilities, like R's `pnorm()`. No explicit integration should be necessary.]

(b) What is the value of this integral?

$$\int_3^4 \int_0^1 e^{-.5x^2+2x-.25y^2+.5y-2.25} dy\, dx$$

[Hint: Again, no explicit integration should be required, and the result from the previous part of this question should get you most of the way to the answer to this one.] You can use the computer to do algebra and integrals if you like, but on this question, if you get the answer by direct numerical integration from a computer, explain how this integration is related to the answer to 1a

(c) What is the joint distribution of $Z_1 = X + Y$ and $Z_2 = X - Y$?

(d) What is $E[Z_2 \mid Z_1 = 4]$?

(2) Estimate a linear regression of `testscr` on `str` and a constant, using the `caschool` data. Also estimate a linear regression of `testscr` on a constant and the following list of variables from that same data set:

```
str
comp_stu
meal_pct
calw_pct
avginc
el_pct
teachers
```

(a) Calculate 95% probability intervals for the coefficient on `str` in the two regression equations. Do they overlap? Can we easily interpret the regression with more variables as simply shrinking our uncertainty about the causal effect of changing the student teacher ratio, compared to what we found with the smaller regression? Or does the larger regression suggest the smaller one was mistaken? [There are enough degrees of freedom here even in the large regression to make using the normal distribution to calculate the probability intervals (instead of the t) a very good approximation. The regression output (e.g., `lmout <- lm()` followed by `summary(lmout)`, in R) should give you standard errors on the estimated coefficients, which you can treat as determining the normal distribution on the coefficients.]

(b) Looking at the the variables that seem to be important in the larger regression, as judged by how clearly the data suggest that they have non-zero coefficients, suggest a substantive interpretation of why the `str` variable's estimated coefficient changes as it does between the two regressions.