Note that the first 2 problems do not require the use of R.

**Problem 1.**

Imagine we have a dataset consisting of companies in the US, and for each company its market value, number of employees, number of female employees, number of male employees, total salary paid, and average salary per employee is recorded. We would like to see if we can find a (linear) relationship between the companies’ market value and employee and salary details.

1. What is the response variable? Market value
2. What predictors do we have?  
   Number of employees, number of female employees, number of male employees, total salary paid, and average salary per employee
3. Say you want to use all the predictors. What would p be equal to? What do you think would be the rank of the model matrix? Will you be successful?  
   . The rank would be 5, since the sum of the female and male employees equals the number of employees. Note that the total salary paid equals the number of employees times the average salary per employee. However, this is not a LINEAR relationship, so it does not affect the rank.
4. Design your analysis: State the model and the predictors you choose to use.  
   Market value

**Problem 2.**

You work for a company that has performed an experiment, recording the value of a variable vs. predictors and They designed the experiment so the predictors are orthogonal to each other. They expect the relationship to be linear and ask you to take the resulting data and estimate the model. You come up with the following:

When you present the data to your superiors, they say you haven’t been listening since you were supposed to only use predictors 3 and 7 (and the intercept).

What will you do?  
Tell them the correct model is:

**Problem 3.**

The data set **prostate** in the faraway library comes from a study on 97 men with prostate cancer who were due to receive a radical prostatectomy. Investigate this data set.

1. First investigate the data. Is there anything you feel that needs to be “fixed” before doing any analysis?  
   svi only has the numbers 0 and 1 in it, so it is a categorical variable. Change it to a factor.
2. Letting *lpsa* be the response, create 2 linear models, one with all the predictors, and one with only *lcavol*, *lweight*, and *svi* as predictors. Call the larger model and the smaller one , and state the two predicted models. Report , , , and .
3. We want to compare the two models. State the hypotheses for the comparison in two different ways (first using the model, second using elements of the parameter vector). Using , assume that the errors of the models are normal and test the two models. Which model will you choose?  
   The p-value so we fail to reject . We will use model .
4. Provide the proportion (or percentage) of improvement of over . 7.58%
5. Now compare the two models again, but this time do NOT assume that the errors are normal. Which model will you choose (this is challenging)?  
   The p-value so we fail to reject . We will use model . Note that the p-value is very close to the normal one.
6. Perform another hypothesis test to see if the parameter for *lcavol* in the model equals .7. Assume normality. State the hypotheses. With , what is your conclusion?  
   p-value so we reject ; we have enough evidence to conclude that the parameter for *lcavol* does not equal .7.
7. Do not assume normality this time and provide a 95% confidence interval for the parameter of *lcavol* in model . Compare it to the CI when normality does hold.  
   Without normality: . With: They are very close.
8. Plot a 95% confidence region (an ellipse) for the parameters for *lweight* and *svi* in , and add the individual confidence intervals. If you wanted to test the null hypothesis that the parameter for lweight=.4 and for svi=.2, what would be your conclusion? What if you just wanted to test for svi=.2?  
   We fail to reject lweight=.4 and svi=.2, since the point is in the ellipse. But we reject svi=.2 since it does not lie in the individual CI.
9. Using the best model you found in b), provide a point estimate for the case where lcavol=1.3, lweight=3, and svi=1. Also provide a 95% PI for a new observation with these values for the predictors.  
   Point estimate=2.64, 95% PI:
10. What would be the 99% CI for the average of future observations with those predictor values?