VARIABLE SELECTION

This exercise concerns the data set SMSA:

```
> smsa = na.omit(read.table("http://pages.uoregon.edu/dlevin/DATA/smsa.txt",
+                           sep="\t",header=T,row.names=1))
```

Set aside 1/3 of the data for testing, use the remaining for training. We want to predict `Mortality` from the other variables.

Omit the variable `NOx` in fitting the largest model, as otherwise the model matrix may be singular. (What does this mean?)

Search for low BIC models using:

- Forward selection
- Backward stepwise selection
- All subset regression

Fit also the following models:

- The ridge regression model with choice of λ minimizing cross-validation error. Choose the number of folds equal to 3 for cross-validation due to small sample size. (Use the option `nfolds` to cv.glmnet)
- The principle component regression models when regressing on 2 and 5 principle components. (Use `ncomp` argument to `predict`.)
- The model with lowest C_p.

Then test these models on the testing data. Find the root-mean-squared-error of prediction for each choice of model. Which models do the best at predicting on the test data?

How sensitive are the recorded RMSE of prediction to the random partitioning into test and training subsets? If results are variable, why is this?

See `http://pages.uoregon.edu/dlevin/DATA/SMSA.html` for description of variables.

Some notes:

- To do forward section using `step` us

  ```
  step(f0,formula(f1),direction="forward",k=log(39))
  ```

  where `f0` is the model with only an intercept, and `f1` is the largest model. Setting `k=\log(39)` uses BIC instead of AIC, where the sample size is 39.
- Use `regsubsets` from the package `leaps` to get all subsets. Use

  ```
  all.mods = regsubsets(Mortality~.-NOx, data=smsatrain, nvmax=17)
  ```

  where `smsatrain` is the training data.