Stepwise Regression

```
require(mosaic)
require(Stat2Data)
```

The book uses the `FirstYearGPA` data set to illustrate automated procedures for variable selection in multiple regression. Since the book uses Minitab, we will work through some examples of how to do this in R.

First, we have to load the data.

```
data(FirstYearGPA)
head(FirstYearGPA)
```

Keep in mind that there is no such thing as a "best" multiple regression model. Each model has its strengths and weaknesses, and it is your job as the data anlayst to construct a model that is effective for your purposes. That could mean forgoing a multiple with a higher R^2 for one that is conceptually simpler. Or it could mean added polynomial terms that explain only a little bit of extra variation in the response.

With this caveat, there is no way to fully automate a procedure that will find the best regression model. Thus, there are several algorithms for optimization a given criterion. There are at least four different optimization procedures:

- Best subset selection
- Backwards elimination
- Forward selection
- Stepwise regression

We will discuss each. Furthermore, each of these techniques can optimize a different criterion. There is no universally agreed-upon best criterion, but the following are popularly used:

- Adjusted R^2
- Mallow's C_p
- Akaike's Information Criterion (AIC)
- Bayesian Information Criterion (BIC)

While the book focuses on Mallow's C_p, we will focus on AIC, since that is what R uses by default.

Best subset regression

The most straightforward method for variable selection is to simply try all of the subsets. Unfortunately, if the number of predictors is k, then you have to try all 2^k subsets! Thus, the number of subsets grows exponentially, which is very fast. [Computer scientists refer to such algorithms as having exponential running times, which are considered extremely inefficient.]

The following code will produce best subsets regression.

```
# install.packages("leaps")
require(leaps)
# Reports the two best models for each number of predictors
best <- regsubsets(GPA ~ ., data=FirstYearGPA, nbest=2)
with(summary(best), data.frame(rsq, adjr2, cp, rss, outmat))
```

1. Which model is the best, choosing based on R^2?
2. Which model is the best choosed based on C_p?
3. Which variables are in the best model of size 4?
4. Run the best model of size 4 and report the AIC value by using the `AIC()` command on your model.

Backward Elimination

Backwards Elimination is a simple algorithm that begins by throwing all of the terms into the model, and then greedily removing the ones that are least statistically significant.

```
backward <- regsubsets(GPA ~ ., data=FirstYearGPA, nbest = 1, nvmax = 6, method = "backward")
summary(backward)
with(summary(backward), data.frame(cp, outmat))
```

1. Which model is the best based on C_p?
2. Find the regression summary for the model with 5 predictors. Are the coefficients significant? Are the assumptions met?
3. What is the AIC value for this model?

Forward Selection

Forward Selection is the opposite idea of Backwards Elimination. Here, we begin with an empty model and then greedily add terms that have a statistically significant effect.

```
forward <- regsubsets(GPA ~ ., data=FirstYearGPA, nbest = 1, nvmax = 6, method = "forward")
with(summary(forward), data.frame(cp, outmat))
```

1. Which model is the best based on C_p?
2. Which variables are in the "best" model?

Stepwise Regression

Stepwise regression combines the ideas of Backwards Elimination and Forward Selection to move in both directions.

```
stepwise <- regsubsets(GPA ~ ., data=FirstYearGPA, nbest = 1, nvmax = 6, method = "seqrep")
with(summary(stepwise), data.frame(cp, outmat))
```

1. What is the model with 4 predictors chosen by this method?
2. What is its AIC value?

Coming to conclusions

Each of these methods will come up with slightly different models. As the analyst, your job is choose which one you think is best, based on your contextual knowledge and understanding of the algorithms. That being said,

1. Which model do you think is best?