Lecture 6

Dimensional reduction – Parsimonious model selection strategies

If you have a lot of possible variables to include in your model, how do you know which variables to include in your model? One? All of them? A subset? If you only include one variable, you might not get the best predictive power out of the available data. If you include too many variables, you might end up overfitting your data: you might be able to predict the data you have very well, but it will be deceptive and you may have unexpectedly poor predictions for new observations. If you have a lot of variables, your model might be unnecessarily complex, with some variables in the model not contributing much to the predictive power. Recall our principle of parsimony: we want the best predictive power for the simplest model necessary. There is a trade-off we want to make between avoiding complexity and the best predictions without overfitting. How do we strike this balance?

There are several strategies we can apply to this problem. We're going to first address some old-school methods that take advantage of some of the things we have already discussed. Then we'll examine some newer methods that take advantage of technology, some penalized regression approaches, and linear algebra techniques. We won't be able to cover all possible methods in this one lecture, but we'll discuss one or two strategies in detail, and then highlight some others that you can implement in R. References for all methods will be included in the reference list at the bottom of the notes for further investigation.

Let's start with the strategy called backward selection. In this approach to regression, we apply all the variables available to us to a model. And then we begin to remove variables one at a time that fail to be significant until we are left with a model that has only significant variables.

Forward selection works similarly, but you begin with just a single variable, and then add variables one at a time, discarding if they are not significant, until you've considered all the possible variables. These methods are somewhat similar, but they don't necessarily produce the same model in the end. The order in which variables get added or subtracted from a model may change which variables end up being retained.

You can also do a mixed model, where you can either add or subtract variables. Perhaps after doing a correlation table. This strategy is less methodical but may allow you to short-cut the full backward or forward selection if you make some wise initial choices.

Let's work through a backward selection approach. We will focus on the statistics tests of the coefficients in our regression analysis. We will omit analyses for outliers and testing of other factors for the sake of having a clean backward selection example, but keep in mind that these other assumptions are still relevant, still apply, and their analysis will still have to be done at some point.

Let's look at a data set on fertility in the MASS package. After loading the data, we can create a model using all the available data in dataset, a full model. The results of a multiple linear regression analysis is shown below.

Call:
lm(formula = Fertility ~ ., data = swiss)

Residuals:

```
    Min      1Q  Median      3Q     Max
-15.2743  -5.2617   0.5032   4.1198  15.3213
```

Coefficients:

	Estimate	Std. Error	t value	Pr(>|t|)
(Intercept)	66.91518	10.70604	6.250	1.91e-07 ***
Agriculture	-0.17211	0.07030	-2.448	0.01873 *
Examination	-0.25801	0.25388	-1.016	0.31546
Education	-0.87094	0.18303	-4.758	2.43e-05 ***
Catholic	0.10412	0.03526	2.953	0.00519 **
Infant.Mortality	1.07705	0.38172	2.822	0.00734 **

Signif. codes: 0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 7.165 on 41 degrees of freedom
Multiple R-squared: 0.7067, Adjusted R-squared: 0.671
F-statistic: 19.76 on 5 and 41 DF, p-value: 5.594e-10

In this case, one of the variables is above the 0.05 significance level, Examination. So we conclude that we don’t have the evidence to think that this variable coefficient is not zero. Therefore we can eliminate it from the model. We would then rerun the model without this variable.

Call:
lm(formula = Fertility ~ . - Examination, data = swiss)

Residuals:

```
    Min      1Q  Median      3Q     Max
-14.6765  -6.0522   0.7514   3.1664  16.1422
```

Coefficients:

	Estimate	Std. Error	t value	Pr(>|t|)
(Intercept)	62.10131	9.60489	6.466	8.49e-08 ***
Agriculture	-0.15462	0.06819	-2.267	0.02857 *
Education	-0.98026	0.14814	-6.617	5.14e-08 ***
Catholic	0.12467	0.02889	4.315	9.50e-05 ***
Infant.Mortality	1.07844	0.38187	2.824	0.00722 **

Signif. codes: 0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 7.168 on 42 degrees of freedom
Multiple R-squared: 0.6993, Adjusted R-squared: 0.6707
F-statistic: 24.42 on 4 and 42 DF, p-value: 1.717e-10

If we are using the significance level of 0.05, then we are done, and we can report our final model. If we wanted to use a 0.01 level of significance, then we could take another step and remove the Agriculture variable as well.

Notice from the R^2 values from the two models, we didn't lose much value here going from 70.67% to just 69.93%, which is less than a percent of explanatory value, and the adjusted R^2 was an even smaller change.

If you have a lot of variables, this can take some time. But you can't necessarily know from the start which strategy will be most effective, whether you will only include one or two variables, or whether you will include most of them. In general, especially if applying a mixed strategy, don't add or subtract more than one variable at a time. It's hard to deal with the possibility of collinearity if you take large leaps like this.

Another strategy which can be implemented by hand, but which is much easier with technology is the best subset strategy. In this strategy you consider all possible combinations of variables. For instance, if you have 5 variables, you would have to look at the full model with all five variables, the five possible models with 4 variables, the 10 models with three variables, the 10 models with 2 variables, the five models with just one variable (and technically also the one model with no variables). You would then compare all the models to choose the one that is the best fit of all these 32 possible models. (In general, if there are k variables, there are 2^k models to consider.) The number of models to test can become quite large as the number of variables increases. Fortunately, there is a package in R that can do this analysis for you and from which you can extract the best model.

The best subset strategy may also produce a model you would not necessarily have found using the forward or backward selection approaches because those approaches doesn't consider every possible model.

Another strategy to avoid overfitting of models, particularly when you have a lot of data is to set aside some of the data you have (generally around 20% of it, though this can vary). Build your initial model, select variables on the 80% that remains, and then test the model on the part you set aside. This allows you a kind of cross validation that permits you see how your model will behave on data it has not seen before. In such cases you will want to compare the errors on the model and on the test set to see if they are similar (this is often done with root mean square error or similar measurements). If the errors are similar, then the model is not overfit. If the errors on the test set are much larger, then this is a sign of overfitting. In the end, you would want to construct your final model on all the available data, but this intermediate testing phase is done to provide insights you can't get when you use all the data to start with.

A strategy for reducing the dimensionality of a model that depends on linear algebra is called principal component analysis (regression) or PCA. This is beyond the linear algebra we have, but essentially it depends on finding the eigenvalues and eigenvectors of our matrix, which is a combination of (possibly) all our variables in certain proportions. We can then select those principal combinations of components that have the biggest impact on our model and drop the rest. A course on data science or advanced linear algebra class will generally cover this approach.

There are also strategies for calculating penalized regression in R in a variety of ways that help restrict the variables included in our model. These include AIC (Akaike information criterion), BIC (Bayesian information criterion), Ridge regression, LASSO (Least Absolute Shrinkage and Selection Operator), Elastic Net, and others. Some of the packages that assist in automating stepwise regression (forward or backward) in R, use AIC or BIC as cut-off criteria. Sources linked below go into these methods in greater

detail. Typically, advanced courses on regression or data mining will cover these methods in greater detail.

Review for Exam #1

The format of the exam will be the same as the exams from last semester. You will have some data to take home and analyze, and then you'll answer some questions about that analysis during class. And then there will be some questions that you will have to answer only in class. For example, I might produce the R analysis or graphs for you and ask you to interpret those things without being able to refer to your notes.

Topics to focus on:

- Pearson correlation interpretation, relationship to coefficient of determination and slope coefficient
- Spearman's and Kendall's Tau interpretation
- Coefficient of determination interpretation
- Interpretation of slope and intercept in the context of a problem
- Be able to conduct model tests, correlation tests, and coefficient hypothesis tests from R output
- Identifying strength of correlation
- Using scatterplots to determine if linear regression is appropriate
- Checking model assumptions using residual plots
- Selection methods (forward, backward, stepwise, best subset, etc.) (be able to use them at home, describe the process in class)
- Outlier detection methods
- The difference between an influential point and an outlier
- Use the normal equations to solve a small regression example
- Use calculus to find covariance of a distribution
- Confidence and prediction intervals

References:

1. https://faculty.ksu.edu.sa/sites/default/files/probability_and_statistics_for_engineering_and_the_sciences.pdf
2. http://www.sthda.com/english/articles/37-model-selection-essentials-in-r/154-stepwise-regression-essentials-in-r/
3. https://towardsdatascience.com/selecting-the-best-predictors-for-linear-regression-in-r-f385bf3d93e9
4. https://www.statology.org/aic-in-r/
5. https://www.statology.org/bic-in-r/
6. https://machinelearningmastery.com/penalized-regression-in-r/