Variable Selection

* A very common problem in regression is to decide on a set of covariates to be included in the model.

* We will generally collect observations on a relatively large number of covariates and then use some of the techniques in this chapter to decide on a "best" model.

* In variable selection there are two, often competing, criteria to be considered:
  1. The selected model should fit the data well;
  2. A simpler model (fewer covariates) is preferred over a more complex model. This is called the Principle of Parsimony.

The Problem

* Suppose that the we have a response variable Y and potential covariates $X_1, \ldots, X_q$.

* The full linear model is

$$Y = \beta_0 + \sum_{j=1}^{q} \beta_j X_j + \varepsilon.$$

* There is a set of covariates $\mathcal{M}_0 \subset \{X_1, \ldots, X_q\}$ such that

$$X_j \in \mathcal{M}_0 \Rightarrow \beta_j \neq 0 \qquad X_j \notin \mathcal{M}_0 \Rightarrow \beta_j = 0$$

* Ideally we would like to use the observed data to find $\mathcal{M}_0$.

Two Important Questions

* Since we are using observed data, we cannot guarantee that we will find the true set $\mathcal{M}_0$.

* Thus there are two important questions that we need to ask:
  1. What is the effect of excluding variables from our selected model which should have been included?
  2. What is the effect of including variables in the model which actually have coefficient equal to zero?

Excluding Important Variables

* Suppose that a variable X_{q+1} should be in the model ($\beta_{q+1} \neq 0$) but we do not include it in the selected model.

* The extra variability in Y which would have been explained by X_{q+1} then becomes part of the error term.

* Thus the estimator of σ^2 will be upwardly biased and the R^2 value will be lower than it need be.

* Furthermore if X_{q+1} is correlated with any variable X_j which is in the model, then the estimator of β_j will be biased and there may be violation of the error assumptions.

Including Extraneous Variables

* Now suppose that the selected model is $\mathcal{M}_0 \cup \{X_{q+1}\}$ but that the true $\beta_{q+1} = 0$.

* The estimators of the coefficients in the model remain unbiased under this scenario.

* However the sampling variances of the estimators are always at least as large as under the true model.

* Since the error degrees of freedom are reduced, the estimator of σ^2 is again biased upwards.

Uses of Regression

* There are 3 main uses of regression modeling
  1. Model building to explain observed variation in the response.
  2. Prediction of future observations.
  3. Control of the magnitude of the change needed in the covariates to effect a given change in the response.

* Although all of these are related, a "best" model for one use may not necessarily be the "best" for another use.

Model Evaluation Criteria

* Given any two models, $\mathcal{M}_1$ and $\mathcal{M}_2$ we need to be able to compare them and say which one is the better model.

* As we said before there are two competing objectives, precise estimation and parsimony.

* For different uses of the regression we may weight these two objectives differently.

* We now look at a few of the model evaluation criteria that have been proposed.

Coefficient of Determination

* Recall that R^2 is the proportion of the variability in Y explained by the model.

* We can use the R^2 value to compare models with the same number of covariates.

* Adding an extra covariate to a model can never reduce R^2 so the best R^2 is always achieved for the full model.

* The adjusted R^2 takes into account the number of covariates. For a model with p covariates

$$R_a^2 = 1 - \frac{\text{SSE}/(n-p-1)}{\text{SST}/(n-1)} = 1 - \frac{n-1}{n-p-1}(1-R^2).$$

* We can compare two models based on their R_a^2 values and select the one with the largest value.

Mallow's C_p

* Introduced by Mallow's in 1973.

* Useful when the primary goal is prediction.

* The standardized total squared error in prediction is

$$J = \frac{1}{\sigma^2} \sum_{i=1}^{n} \mathsf{E}\left((y_i - \widehat{y}_i)^2\right).$$

* For a model $\mathcal{M}$ with p covariates estimator of J is

$$C_p = \frac{\mathsf{SSE}(\mathcal{M})}{\widehat{\sigma}^2} + 2(p+1) - n$$

where $\widehat{\sigma}^2$ is the estimator of σ^2 obtained from the full model.

Information Criteria

* Based on log-likelihood ratios for comparing two models.
* Also include a penalty for the number of parameters.
* Akaike's Information Criteron (AIC)

$$\text{AIC}(\mathcal{M}) \;=\; n\log(\text{SSE}(\mathcal{M})/n) + 2(p+1)$$

* Bayes Information Criterion (BIC)

$$\text{BIC}(\mathcal{M}) \;=\; n\log(\text{SSE}(\mathcal{M})/n) + (p+1)\log(n)$$

Information Criteria

* Models with smaller values of the criteria are preferred.

* For models with the same number of parameters both criteria will select the one with the smallest SSE.

* For $n > 7$, the penalty in BIC is greater than that in AIC and so tends to result in models with smaller numbers of parameters.

All Subsets Regression

* Consider all possible subsets of $\{X_1, \ldots, X_q\}$.
* Fit a regression for each one.
* Compare the fits using one of the criteria mentioned above.
* For large q, this is very computationally expensive since it requires 2^q regressions!
* As n gets large, this method will almost always select the correct model.

Testing Based Procedures

* It is attractive to use t and F tests in looking for models.
* These, are only valid, however, for nested models.
* Testing based procedures therefore follow a path through the possible subsets such that the model at step r and each of the models examined at step $r+1$ are nested.
* We can then use tests to decide what to do next.
* We can rank the models examined at step $r+1$ based on their test statistics.

Forward Selection

* Starts with the null model (no covariates).
* At each step looks for a single variable not currently in the model to add.
* Any added variable needs to improve the model significantly based on the t or F test.
* If multiple variables are significant then add the one which most improves the model, that is the variable with the largest t statistic in absolute value.

Forward Selection

* Stop when none of the remaining variables significantly improve the fit.

* Forward selection requires fitting at most $1 + q(q+1)/2$ regressions.

* There is no guarantee that the true model will be one of those tried.

* The models tested may not even be the "best" based on any of the criteria mentioned earlier.

Backward Elimination

* Starts with the full model (all q covariates).
* At each stage, tries to remove one of the variables in the model.
* Remove the variable with the smallest t in absolute value provided it is not significant.
* Stop when all variables in the model are significant or no variables remain.
* Backward elimination requires at most $q + 1$ regressions.

Stepwise Regression

* As with Forward Selection, start with the null model and add a variable.

* At subsequent stages, however, we will also consider the possibility of dropping one of the variables before adding another.

* If a variable in the current model is insignificant then delete it, otherwise consider adding a variable.

* May require more fits than forward selection but generally results in a simpler model.

Variability of Estimators

* When we use model selection criteria, we are using the data to select the model.

* For different datasets from the same true model, we would likely end up with different models.

* This introduces extra variability into the estimators of the parameters which is not accounted for in the usual standard errors.

* Nevertheless, it is common to treat the fitted model as if we always planned to fit this model!

* Advances in computational techniques have produced methods for including model uncertainty into the standard errors but these are beyond the scope of this course.

Other Comments

* It is important to remember that the standard regression diagnostics should be run on any potential model.

* Models which violate any standard assumptions should not be considered even if they come up in a selection procedure.

* For stepwise procedures it is common to calculate some (or all) of the criteria mentioned earlier on the models chosen at each step. It is then possible to compare these models.

* In backward elimination, it is also common to examine the F test comparing the final model with the full model.

* If this is significant, then we may choose one of the earlier (larger) models instead.

THE END!!!!!