18: MODEL SELECTION

It is important to remember that the multiple linear regression model is, at best, just an *approximation* to the truth. It will almost never hold exactly. Even if it did hold exactly for some set of explanatory variables, we can never be sure that we have found the right set. In **model selection,** the idea is to find the smallest set of variables which provides an adequate description of the data.

Why isn’t it a good idea to just put every variable we can think of into the model? After all, shouldn’t we use all the "information" we have? Unfortunately, some of this information is highly

redundant. In the housing example, it would be a mistake to include *both* the house size in square feet and the house size in square meters, as the second variable gives us no new information whatsoever. Furthermore, it might not be a good idea to include both the house size and the lot size, especially if these variables are deemed to be highly correlated.

Another problem is that even pure "noise" will start to look like signal if we examine it carefully enough. For example, suppose we regress the price of gold (y) on the price of silver (x_1), the world's whale population (x_2), and the number of rainy

days in New York City (x_3), for each of the past 25 years. Presumably, x_2 and x_3 are of no use in predicting y. However, a model with all three variables will produce a larger R^2 than the model with just x_1. (Why?) This is one reason why it is often ridiculous to think of R^2 as measuring the proportion of "explained" variation. It is clear, then, that the model with the largest R^2 may not be the best model to use.

Would it really hurt us to have the "garbage" variables x_2 and x_3 in the model above? The answer is often yes, because the coefficient of x_1 may be adversely affected, and also because our

ability to predict y may actually *deteriorate* if we include the extraneous variables. With regard to x_1, there is another issue. Some theories hold that the prices of gold and silver are *inversely* related. Therefore, maybe we should use the *reciprocal* of the silver price instead of x_1. The choice of the appropriate transformation of the explanatory variables is also part of the problem of model selection.

All explanatory variables that are available to us are called **candidate variables.** Some of these variables may be transformations of others. For example, we could have $x_1 = SP$, x_2 and x_3 as above, $x_4 = 1/SP$, $x_5 = \log(SP)$, where SP is the price of

silver. We want to decide which *subset* of these candidate variables should be used in our multiple regression model. Unfortunately, if there are K candidate variables, then there are 2^K possible subset models. This grows very fast with K, so we may not be able to look at all subsets. The idea of *stepwise methods* (see Jobson, Section 4.2.1) is to avoid looking at all subsets. Whether or not we have the computing power, and energy, to look at all subsets, we will need some measure of the quality of a given *candidate model* (i.e., using a particular subset of the candidate variables). We have already seen that R^2 is not an adequate measure,

essentially because it does not penalize the model for having too many parameters. Neither does the overall F-statistic.

Using the t-ratios for model selection can lead to contradictory results, depending on which candidate model is used. For the housing example, age was *not* significant ($p = .8$) in the full model, but *was* significant ($p = .047$) when it was the only variable used.

For each candidate model, we would like to construct a numerical measure of quality which is directly comparable to the quality measures obtained for the other models, which automatically takes into

account the fact that different models will often have different numbers of explanatory variables, and which automatically trades off our conflicting objectives to get a reasonably good description of the data (e.g., make *SSE* small), without using too many variables. We will consider two different **information criteria** for this purpose. The criteria are called *FPE* (Final Prediction Error), and AIC_C (Corrected Akaike Information Criterion). Both involve evaluating the criterion for the various candidate models, and then choosing the model which *minimizes* the criterion.

The FPE Criterion

Consider the candidate model $y = X\beta + u$, where X is $n \times (p+1)$ and the u_i are *iid* with mean zero and variance σ_u^2. We do not need to assume that the u_i are normal. Next, we *suppose that this candidate model is correct.* This assumption will usually be wrong, but it allows us to derive our selection criterion! Using the data y, we obtain the predicted values, $\hat{y}$.

Imagine a vector of future observations, $y^F = X\beta + u^F$, where the u_i^F are *iid* with mean zero, variance σ_u^2, and are *independent* of the u_i. The sum of squared prediction errors corresponding

to the predictor $\hat{y}$ is

$$PE = ||y^F - \hat{y}||^2 = ||X\beta + u^F - Xb||^2$$

$$= ||X(\beta - b) + u^F||^2$$

$$= ||X(\beta - b)||^2 + ||u^F||^2 + 2(u^F)'X(\beta - b) \ .$$

The expectation of PE is called the **Final Prediction Error.** Since since u^F and b are independent, the final prediction error is given by

$$E[PE] = E[(\beta - b)'X'X(\beta - b)] + n\sigma_u^2$$

$$= \sigma_u^2(p+1) + n\sigma_u^2 = \sigma_u^2(n+p+1) \ ,$$

which can be estimated without bias by

$$FPE = s^2(n+p+1) = SSE \ \frac{n+p+1}{n-p-1} \ .$$

Note that *FPE* automatically compromises the fidelity to the data and "parsimony". The fidelity is measured by *SSE*, which tends to decrease with p. The other term, $(n+p+1)/(n-p-1)$, which increases with p, constitutes a penalty term designed to prevent us from using too many parameters.

To implement this method, we get the residual sum of squares *SSE* for each candidate model, and then calculate *FPE*. We select the model which gives the smallest value of *FPE*.

The Corrected Akaike Information Criterion, AIC_C

Unfortunately, in small samples, or whenever the total number of candidate models is large, *FPE* will

often select too many variables. Another method, which works better in small samples (and just as well as *FPE* in large samples) is the corrected Akaike Information Criterion,

$$AIC_C = \log(SSE) + \frac{2(p+2)}{n-p-3} \quad ,$$

where log denotes the *natural* logarithm. The model selected by minimizing AIC_C will often have fewer variables than the one selected by *FPE*. To see why, note that the model which minimizes *FPE* will also minimize *log* (*FPE*), given by

$$\log(FPE) = \log(SSE) + \log((n+p+1)/(n-p-1)) \quad .$$

Some calculus reveals that the penalty term for AIC_C increases much faster as a function of p than

the penalty term for $log(FPE)$, particularly when p is large.

An Example

We consider a data set on the utilization of trees for the production of matches. For each of 1790 trees, the "expected utilization" was measured (as a percentage of the volume), along with the diameter (in inches) at breast height. The data were then grouped according to 16 classes, according to diameter. Our 16 data points consist of x, the diameter representing the class, and y, the average utilization for the trees in the given class. We want to understand how utilization depends on diameter.

A scatterplot reveals that the relationship between x and y is not linear. Although utilization increases at first, it seems to eventually level off.

We will try fitting polynomials of degrees 1-7 to the data. To fit a polynomial of degree p, we simply use the explanatory variables $x_1, \ldots, x_p$, where $x_j = x^j$. (Of course, we also include x_0, a column of ones.) The resulting polynomial regression model is

$$y = \beta_0 + \beta_1 x + \cdots + \beta_p x^p + u \quad .$$

This is a linear regression model, since it is **linear in the parameters,** even though $E[y|x]$ is not a linear function of the diameter, x.

We will also try an *exponential* model, of form

$$y = \alpha + \beta \rho^x + u \quad ,$$

where α, β, ρ are unknown parameters. This model is **nonlinear,** since $E[y]$ cannot be expressed as a linear function of the parameters. It is still possible to fit the model by least squares, but we need to use a nonlinear optimization package to find the parameter values which minimize the sum of squares criterion function. The resulting fitted model is

$$\hat{y} = 93.72 - 2234\,(.666)^x \quad .$$

This model seems to fit the data well, using only 3 parameters. As x increases, the predicted utilization levels off at 93.72.

It could be argued that the true regression function $E[y|x]$ should increase with x but eventually level off. (It certainly cannot exceed 100.) No polynomial model could produce this behavior. In principle, then, the exponential model seems most appropriate. Now, let's see what conclusions are reached by FPE and AIC_C, which of course know nothing of the above considerations.

Table I gives p, SSE, R^2, FPE and AIC_C for the 8 candidate models. For the polynomial models, SSE decreases and R^2 increases with p, as expected, FPE selects a 6'th degree polynomial, and AIC_C selects a 4'th degree polynomial.

Plots of the FPE and AIC_C criteria, as functions of p, reveal that AIC_C exhibits a clear minimum at $p = 4$, while FPE seems to be wandering generally downward. (In fact, still higher values of p, not shown here, would actually be preferred by FPE.)

If the exponential model is now included as a candidate, AIC_C selects it as the overall best model, in accordance with our intuition above, while FPE ranks it as only fifth best.

Data Splitting

A problem with model selection as described above is that the usual inferential procedures based on the selected model will not be completely valid. The process of selecting a model entails a certain amount of "data snooping". In a rough sense, we are picking out the variables that seem to be the most significant. So it would be wrong to pretend that this snooping never happened. For example, given that x_1 is in the selected model, the usual t-test of $H_0 : \beta_1 = 0$ will have a type I error rate which is much larger than the nominal level, α. (Why?)

A way around the problem is **data splitting.**

Split the data into two subsamples, A and B, selected at random, of sizes n_1 and n_2. Use A for model selection, without examining B. Then, using this selected model, re-estimate the parameters for the data in B. If the selected model is correct, then all inferences based on the data in B will be valid. This procedure may seem wasteful (why?). Fortunately, it is not necessary to use very much of the available data for the model selection stage. For example, taking n_1 to be 10% of the overall sample size n should be adequate, assuming that AIC_C is used, and $n \geq 100$.