Part 6: Model Selection and Intro to ML

Chris Conlon

March 6, 2021

Applied Econometrics II

Stepwise Regression

Back to the real world...

- We have some theoretical benchmark which lets us discern which of two model we prefer (under certain assumptions).
- In practice we often start with a functional form like:
  $y_i = \beta_0 + \sum_{k=1}^{p} \beta_k x_{i,k} + \varepsilon_i$
- Which x's do we include?
- Which x's do we leave out?
- It is not clear that BIC/AIC or Vuong test tells us what we should do in practice.
- If you have K potential regressors you could consider all 2^K possible regressions.
- Or you could could consider all $\binom{K}{P}$ possible combinations with p parameters.
- This sounds very time consuming

Things to keep in mind

Two major (related) problems:

- Regressors are correlated with one another:
  - small changes in the sample: β_1 goes up, β_2 goes down.
  - large coefficients can lead to wild predictions.
  - If relationship between y_i and x_i is nonlinear, and (x_i, z_i) are highly correlated then we may attribute some of this nonlinearity to z_i, even when it has no effect.
- Lots of imprecisely estimated parameters can make prediction tricky
  - Small changes in the sample can lead to large changes in $\hat{y}_i|x_i$.

The big idea: maybe we tolerate some bias to greatly reduce variance.

- This is where ML and Econometrics diverge!
- Econometrics historically focuses on unbiasedness.

Minimizing SSR/AIC: all possible regressions

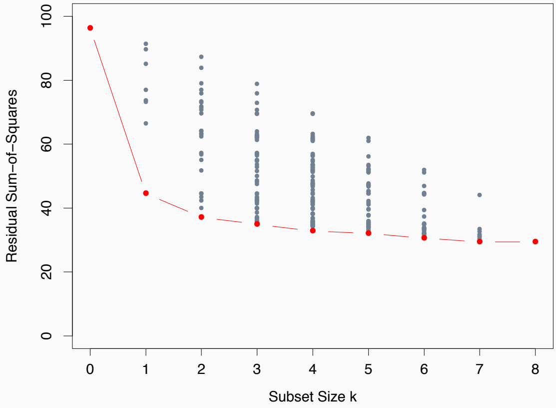

FIGURE 3.5. *All possible subset models for the prostate cancer example. At each subset size is shown the residual sum-of-squares for each model of that size.*

What is orthogonality?

- We can think about a world where $\langle x_j, x_k \rangle = 0$ for $j \neq k$.
- In this world I can get β_j by regressing y on x_j by simple linear regression.
- I could do this for each j and the resulting vector β would be the same as running multiple regression.
- We could try and transform X so that it forms an orthogonal basis.
- Unless we are running regressions by hand this doesn't seem tremendously helpful.
- However, in practice this is often what your software does!

Gram-Schmidt/QR Decomposition

1. Let $x_0 = z_0 = 1$
2. For $j = 1, 2, \dots p$: Regress x_j on $z_0, z_1, \dots, z_{j-1}$ to give you $\hat{\gamma_{jl}} = \langle z_l, x_j \rangle / \langle z_l, x_l \rangle$ and residual $z_j = x_j - \sum_{k=0}^{j-1} \hat{\gamma_{kj}} z_k$.
3. With your transformed orthogonal basis $\mathbf{z}$ you can now regress y on z_p one by one to obtain $\hat{\beta}_p$.

What does this do?

- The resulting vector $\hat{\beta}$ has been adjusted to deliver the marginal contribution of x_j on y after adjusting for all x_{-j}.
- If x_j is highly correlated with other x_k's then the residual z_j will be close to zero and the coefficient will be unstable.
- This will be true for any variables x_l within a set of correlated variables.
- We can delete any one of them to resolve this issue.

QR Decomposition (Technical Details)

QR Decomposition has a matrix form which regression software uses:

$$
\begin{aligned}
\mathbf{X} &= \mathbf{Z\Gamma} \\
&= \underbrace{ZD^{-1}}_{\mathbf{Q}}\underbrace{D\Gamma}_{\mathbf{R}} \\
\hat{\beta} &= \mathbf{R}^{-1}\mathbf{Q}'\mathbf{y} \\
\hat{\mathbf{y}} &= \mathbf{QQ}'\mathbf{y}
\end{aligned}
$$

- Z is the matrix of the orthogonalized residuals z_j's.
- Γ is upper triangular matrix with entries $\hat{\gamma}_{kj}$
- D is diagonal matrix with entries $||z_j||$.
- Q is $N \times ((p+1)$ orthogonal matrix $Q'Q = I$
- R is $(p+1) \times (p+1)$ upper triangular matrix.

What happens in practice?

What are people likely doing in practice:

- Start with a single x variable and then slowly add more until additional x's were insignificant
- Start with all possible x variables and drop those where t-statistics were insignificant.
- These procedures actually make some sense if the columns of X are linearly independent or orthogonal.
- In practice our regressors are often correlated (sometimes highly so).

Forward Stepwise Regression

Consider the following greedy algorithm

1. Start with an empty model and add a constant $\overline{y}$.
2. Then run K single-variable regressions, choose the x_k with the highest t-statistic call this $x^{(1)}$.
3. Now run $K-1$ two variable regressions where the constant and $x^{(1)}$ and choose $x^{(2)}$ as regression where x_k has the highest t-statistic.
4. Now run $K-2$ three variable regressions where the constant and $x^{(1)}, x^{(2)}$
5. You get the idea!

We stop when the x_k with the highest t-statistic is below some threshold (often 20% significance).

Backwards Stepwise Regression

1. Start with an full model.
2. Remove the x variable with the lowest t-statistic. Call this $x^{(k)}$.
3. Re-run the regression without $x^{(k)}$.
4. Repeat until the smallest t-statistic exceeds some threshold.

Comparison

- Backwards and fowards stepwise regression tend to give similar choices (but not always).
- Everything is trivial if X's columns are orthogonal (computer has some tricks otherwise- QR).
- Forward stepwise works when we have more regressors than observations $K > N$.
- I proposed the t-stat here but some packages use AIC/BIC as the criteria.
- We should also be careful to group dummy variables together as a single regressor.
- These are implemented in `step` in R and `stepwise` in Stata.
- We probably want to adjust our standard errors for the fact that we have run many regressions in sequence before arriving at our model. In practice not enough people do this!

(Incremental) Forward Stagewise Regression

As an alternative consider:

1. Start with $r = y$ and $(\beta_1, \ldots, \beta_p) = 0$.
2. Find the predictor x_j most correlated with r.
3. Update $\beta_j \leftarrow \beta_j + \delta_j$ where $\delta_j = \epsilon \cdot sgn\langle r, x_j \rangle$.
4. Update $r \leftarrow r - \delta_j \cdot x_j$ and repeat for S steps.

- Alternative $\delta_j = \langle r, x_j \rangle$
- We can continue until no regressors have correlation with residuals
- This is very slow (it takes many many S).
- Sometimes slowness can be good – in high dimensions to avoid overfitting.

Stepwise selection proedures

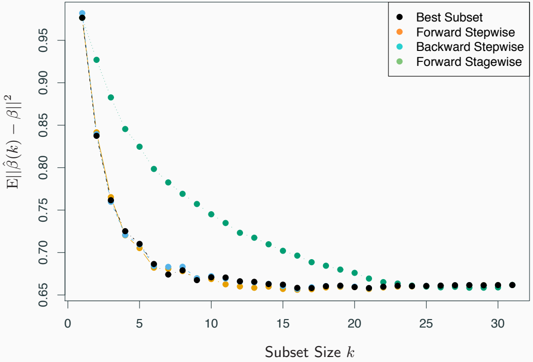

FIGURE 3.6. *Comparison of four subset-selection techniques on a simulated linear regression problem $Y = X^T\beta + \varepsilon$. There are $N = 300$ observations on $p = 31$ standard Gaussian variables, with pairwise correlations all equal to 0.85. For 10 of the variables, the coefficients are drawn at random from a $N(0, 0.4)$ distribution; the rest are zero. The noise $\varepsilon \sim N(0, 6.25)$, resulting in a signal-to-noise ratio of 0.64. Results are averaged over 50 simulations. Shown is the mean-squared error of the estimated coefficient $\hat{\beta}(k)$ at each step from the true β.*

Multiple Testing Problem

- A big deal in Econometrics frequently ignored in applied work is the Multiple Testing Problem
- You didn't just pick the regression in your table and run that without considering any others.
- This means that your t and F stats are going to be too large!! (Standard errors too small!)
- How much bigger should they be?
    - Analytic size corrections can be tricky and data dependent
    - Bootstrap/Monte-Carlo studies should give you a better idea.