Day 4: Shrinkage Estimators

Kenneth Benoit

Data Mining and Statistical Learning

March 9, 2015

n versus p (aka k)

- Classical regression framework: $n > p$. Without this inequality, the OLS coefficients have no unique solution
- The variance of the estimates increases as $p \rightarrow n$
- To predict problems where $n < p$, we need new strategies - OLS and versions of OLS will not work
- Note: These are predictive methods, not methods for explanation!

Strategies for coping with $n < p$

- Subset selection. Identifying a relevant subset of the $p < n$ predictors, and fitting an OLS model on the reduced set of variables
- Shrinkage. Fitting a model involving all p predictors, but penalizing (regularizing) the coefficients in such as way that they are shrunken towards zero relative to the least squares estimates
  - has the effect of reducing variance
  - may also perform variable selection (with the lasso)
- Dimension Reduction. Replacing the p predictors with projections (linear combinations) of the predictors onto M-dimensional subspace, where $M < p$, and then fitting an OLS model on the reduced set of (combination) variables

Subset selection through sequential testing

- Such a model can be found using a series of significance tests
  - Usual t or F tests of the coefficients, all using the same significance level (e.g. 5%)
- Two basic versions are:
  - Forward selection: start with a model with no explanatory variables, and add new ones one at a time, until none of the omitted ones are significant
  - Backward selection: start with a model with all the variables included, and remove nonsignificant ones, one at a time, until the remaining ones are significant
- In practice, the most careful (and safest) procedure is a combination of these two: stepwise selection

Stepwise model selection

- Pick a significance level, say $\alpha = 0.05$
- Start forwards (from a model with nothing in it) or backwards (from a model with everything in it)
- Then, repeat the following:
  - Add to the current model the omitted variable for which the P-value would be smallest, if this is smaller than α
  - Remove the variable with the largest P-value, if this is bigger than α
  - Continue until all the variables in the current model have $P < \alpha$, and all the ones out of it would have $P \geq \alpha$
- This process can even be done "automatically" in one go, but usually shouldn't — needs more care than that

Example from HIE data

- Response variable: General Health Index at entry, $n = 1113$
- Potential explanatory variables: sex (dummy for men), age, log of family income, weight, blood pressure and smoking (as two dummy variables, for current and ex smokers)
  - A haphazard collection of variables with no theoretical motivation, purely for illustration of the stepwise procedure
  - For simplicity, no interactions or nonlinear effects considered
- F-tests are used for the smoking variable (with two dummies), t-tests for the rest
- Start backwards, i.e. from a full model with all candidate variables included

Example: Health Index Experiment

Variable	Response variable: General Health Index Model (1)	(2)	(4)	(5)	(6)
Age	−0.138 (< 0.001)	−0.089 (0.004)	−0.128 (< 0.001)	−0.142 (< 0.001)	—
Education	—	1.157 (< 0.001)	0.990 (< 0.001)	0.981 (< 0.001)	1.117 (< 0.001)
Income	—	—	0.275 (< 0.001)	0.277 (< 0.001)	0.219 (< 0.001)
Work experience	—	—	—	0.002 (0.563)	−0.007 (0.045)
(Constant)	74.777	58.801	59.417	59.723	54.666
R^2	0.012	0.051	0.061	0.061	0.054

(P-values in parentheses)

Example from HIE data

1. In the full model, Blood pressure ($P = 0.71$), Smoking ($P = 0.30$) and Sex ($P = 0.19$) are not significant at the 5% level
   - Remove Blood pressure
2. Now Smoking ($P = 0.30$) and Sex ($P = 0.19$) are not significant
   - Remove Smoking
3. If added to this model, Blood pressure is not be significant ($P = 0.71$), so it can stay out
4. In this model, Sex ($P = 0.21$) is the only nonsignificant variable, so remove it
5. Added (one at a time) to this model, neither Blood pressure ($P = 0.77$) nor Smoking ($P = 0.31$) is significant, so they can stay out

Example from HIE data

- So the final model includes Age, Log-income and Weight, all of which are significant at the 5% level
- Here the nonsignificant variables were clear and unchanging throughout, but this is definitely not always the case

Comments and caveats on stepwise model selection

- Often some variables are central to the research hypothesis, and treated differently from other control variables
  - e.g. in the Health Insurance Experiment, the insurance plan was the variable of main interest
  - Such variables are not dropped during a stepwise search, but tested separately at the end
- Variables are added or removed one at a time, not several at once
  - For categorical variables with more than two categories, this means adding or dropping all the corresponding dummy variables at once
  - Individual dummy variables (i.e. differences between particular categories) may be tested separately (e.g. at the end)

Comments and caveats on stepwise model selection

- The models should always be hierarchical:
  - if an interaction (e.g. coefficient of X_1X_2) is significant, main effects (X_1 and X_2) may not be dropped
  - if coefficient of X^2 is significant, X may not be dropped
- In practice, the possible interactions and nonlinear terms are often not all considered in model selection
  - Only those with some a priori plausibility
- Because it involves a sequence of multiple tests, the overall stepwise procedure is not a significance test with significance level α
- Not guaranteed to find a single "best" model, because it may not exist: there may be several models satisfying the conditions stated earlier

Changes of scale

- In short: Linear rescaling of variables will not change the essential key statistics for inference, just their scale
- Suppose we reexpress x_i as $(x_i + a)/b$. Then:
  - t, F, $\hat{\sigma}^2$, R^2 unchanged
  - $\hat{\beta}_i \rightarrow b\hat{\beta}_i$
- Suppose we rescale y_i as $(y_i + a)/b$. Then:
  - t, F, R^2 unchanged
  - $\hat{\sigma}^2$ and $\hat{\beta}_i$ will be rescaled by b
- Standardized variables and standardized coefficients: where we replace the variables (all x and y) by their standardized values $(x_i - \bar{X})/\text{SD}_x$ (e.g. for x). Standardized coefficients are sometimes called "betas".

More on standardized coefficients

Consider a standardized coefficient b^* on a single variable x.

- Formula: $b^* = b\frac{\text{SD}_x}{\text{SD}_y}$
- Intrepretation: the increase in standard deviations of y associated with a one standard deviation increase in x
- Motivation: "standardizes" units so we can compare the magnitude of different variables' effects
- In practice: serious people never use these and you should not either
  - too tricky to interpret
  - misleading since suggests we can compare apples and oranges
  - too dependent on sample variation (just another version of R^2)
- We can illustrate this in R, if we use the `scale()[,1]` command to standardize the variables, which transforms them into $z_i = (x_i - \bar{X})/\text{SD}_x$

Standardized coefficients illustrated

```
> dc <- d[complete.cases(d$votes1st, d$spend_total),] # remove missing
> m1.std <- lm(scale(dc$votes1st)[,1] ~ scale(dc$spend_total)[,1])
> coef(m1)
(Intercept) spend_total
683.7550298   0.2336056
> coef(m1.std)
               (Intercept) scale(dc$spend_total)[, 1]
             -1.100854e-16                7.395996e-01
> coef(m1)[2]*sd(dc$spend_total)/sd(dc$votes1st)
spend_total
  0.7395996
```

Collinearity

- When some variables are exact linear combinations of others then we have exact collinearity, and there is no unque least squares estimate of β
- When X variables are correlated, then we have (multi)collinearity
- Detecting (multi)collinearity:
  - look at correlation matrix of predictors for *pairwise* correlations
  - regress x_k on all other predictors to produce R_k^2, and look for high values (close to 1.0)
  - Examine eigenvalues of $X'X$

Collinearity continued

- Define:

$$S_{x_j x_j} = \sum_i (x_{ij} - \bar{x}_j)^2$$

then

$$\text{Var}(\hat{\beta}_j) = \sigma^2 \left(\frac{1}{1 - R_j^2} \right) \frac{1}{S_{x_j x_j}}$$

- So collinearity's main consequence is to reduce the efficiency of our estimates of β
- So if x_j does not vary much, then $\text{Var}(\hat{\beta}_j)$ will be large – and we can maximize $S_{x_j x_j}$ by spreading X as much as possible
- We call this factor $\frac{1}{1-R_j^2}$ a variance inflaction factor (the `faraway` package for R has a function called `vif()` you can use to compute it)
- *Orthogonality* means that variance is minimized when $R_j^2 = 0$

Model fit: Revisiting the OLS formulas

For the three parameters (simple regression):

- the regression coefficient:

$$\hat{\beta}_1 = \frac{\sum(x_i - \bar{x})(y_i - \bar{y})}{\sum(x_i - \bar{x})^2}$$

- the intercept:

$$\hat{\beta}_0 = \bar{y} - \hat{\beta}_1 \bar{x}$$

- and the residual variance σ^2:

$$\hat{\sigma}^2 = \frac{1}{n-2} \sum [y_i - (\hat{\beta}_0 + \hat{\beta}_1 x_i)]^2$$

OLS formulas continued

Things to note:

- the prediction line is $\hat{y} = \hat{\beta}_0 + \hat{\beta}_1 x$
- the value $\hat{y}_i = \hat{\beta}_0 + \hat{\beta}_1 x_i$ is the predicted value for x_i
- the residual is $e_i = y_i - \hat{y}_i$
- The residual sum of squares (RSS) $= \sum_i e_i^2$
- The estimate for σ^2 is the same as

$$\hat{\sigma}^2 = \text{RSS}/(n-2)$$

Components of least squares model fit

TSS Total sum of squares $\sum(y_i - \bar{y})^2$

ESS Estimation or Regression sum of squares $\sum(\hat{y}_i - \bar{y})^2$

RSS Residual sum of squares $\sum e_i^2 = \sum(\hat{y}_i - y_i)^2$

The key to remember is that TSS $=$ ESS $+$ RSS

R^2

How much of the variance did we explain?

$$R^2 = 1 - \frac{RSS}{TSS} = 1 - \frac{\sum_{i=1}^{N}(y_i - \hat{y}_i)^2}{\sum_{i=1}^{N}(y_i - \bar{y})^2} = \frac{\sum_{i=1}^{N}(\hat{y}_i - \bar{y})^2}{\sum_{i=1}^{N}(y_i - \bar{y})^2}$$

Can be interpreted as the *proportion of total variance explained by the model*.

R^2

- A much over-used statistic: it may not be what we are interested in at all
- Interpretation: the proportion of the variation in y that is explained lineraly by the independent variables
- Defined in terms of sums of squares:

$$\begin{aligned} R^2 &= \frac{ESS}{TSS} \\ &= 1 - \frac{RSS}{TSS} \\ &= 1 - \frac{\sum(y_i - \hat{y}_i)^2}{\sum(y_i - \bar{y})^2} \end{aligned}$$

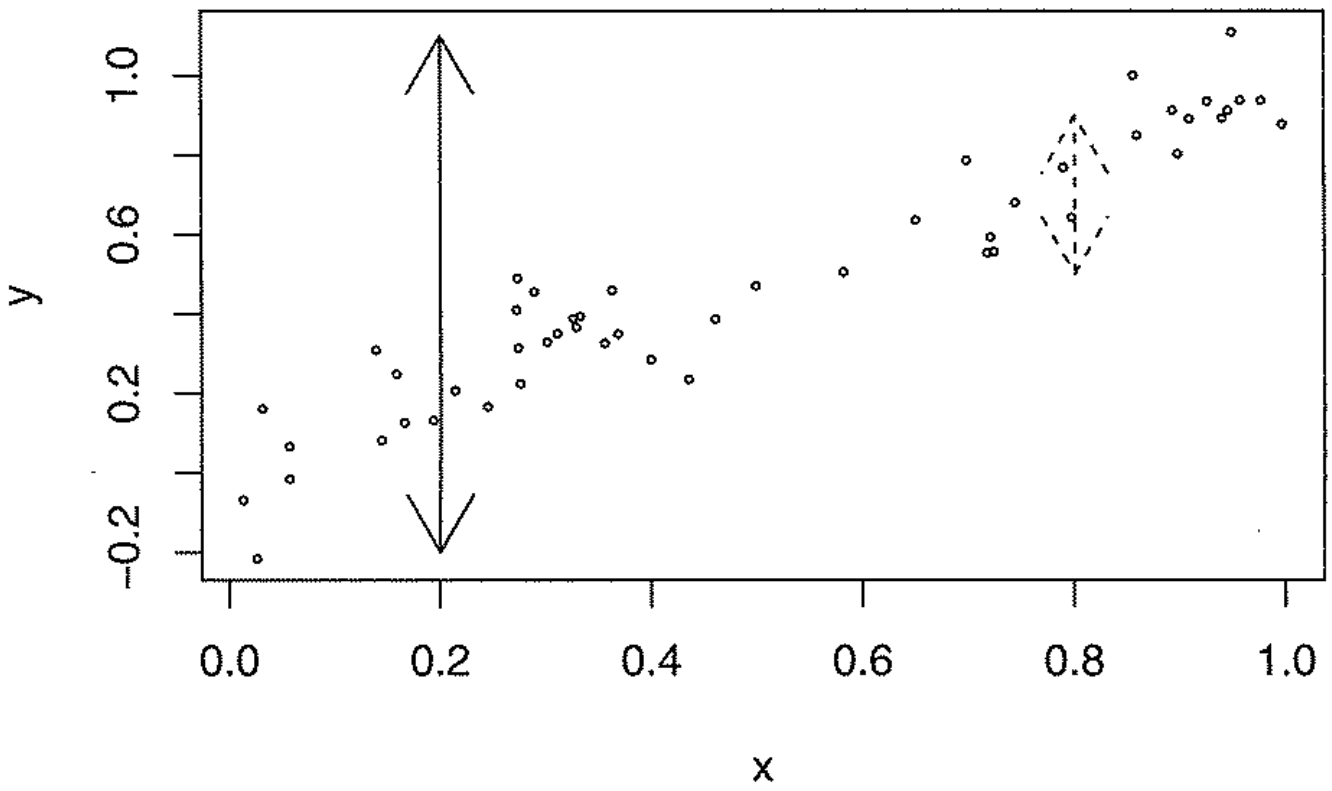
- Alternatively, R^2 is the squared correlation coefficient between y and $\hat{y}$

R^2 continued

- When a model has no intercept, it is possible for R^2 to lie outside the interval $(0, 1)$
- R^2 rises with the addition of more explanatory variables. For this reason we often report "adjusted R^2": $1 - (1 - R^2)\frac{n-1}{n-k-1}$ where k is the total number of regressors in the linear model (excluding the constant)
- Whether R^2 is *high* or not depends a lot on the overall variance in Y
- To R^2 values from different Y samples *cannot be compared*

R^2 continued

$$R^2 = \frac{\sum(\hat{y}_i - \bar{y})^2}{\sum(y_i - \bar{y})^2}$$



- Solid arrow: variation in y when X is unknown (TSS Total Sum of Squares $\sum(y_i - \bar{y})^2$)
- Dashed arrow: variation in y when X is known (ESS Estimation Sum of Squares $\sum(\hat{y}_i - \bar{y})^2$)

R^2 decomposed

$$
\begin{aligned}
y &= \hat{y} + \epsilon \\
\mathrm{Var}(y) &= \mathrm{Var}(\hat{y}) + \mathrm{Var}(e) + 2\mathrm{Cov}(\hat{y}, e) \\
\mathrm{Var}(y) &= \mathrm{Var}(\hat{y}) + \mathrm{Var}(e) + 0 \\
\sum (y_i - \bar{y})^2/N &= \sum (\hat{y}_i - \bar{\hat{y}})^2/N + \sum (e_i - \hat{e})^2/N \\
\sum (y_i - \bar{y})^2 &= \sum (\hat{y}_i - \bar{\hat{y}})^2 + \sum (e_i - \hat{e})^2 \\
\sum (y_i - \bar{y})^2 &= \sum (\hat{y}_i - \bar{\hat{y}})^2 + \sum e_i^2 \\
TSS &= ESS + RSS \\
TSS/TSS &= ESS/TSS + RSS/TSS \\
1 &= R^2 + \text{unexplained variance}
\end{aligned}
$$

Other model fit statistics

Where d is the number of predictors and $\hat{\sigma}^2$ is the estimated residual error variance,

- Mallows's C_p

$$C_p = \frac{1}{n}(RSS + 2d\hat{\sigma}^2)$$

- Akaike information criterion (AIC)

$$\text{AIC} = \frac{1}{n\hat{\sigma}^2}(RSS + 2d\hat{\sigma}^2)$$

(note: perfectly correlated with C_p for OLS)

- BIC

$$\text{BIC} = \frac{1}{n}(RSS + log(n)d\hat{\sigma}^2)$$

penalizes the number of parameters (d) more than AIC

- Adjusted R^2

$$1 - (1 - R^2)\frac{n-1}{n-d-1}$$

Penalized regression

- Provides a way to shrink the variance of estimators (toward zero), to reduce the variance inflation problem that occurs as $p \to n$
- Also solves non-uniqueness of β estimates when $n < p$
- Some methods (e.g. lasso) even shrink estimates to zero, performing a type of variable selection
- Two most common methods:
  - ridge regression
  - lasso regression
- Both involve a "tuning parameter" λ whose value must be set based on optimizing some criterion (usually, predictive fit)

Ridge regression

- OLS: minimize the residual sum of squares, defined as:

$$RSS = \sum_{i=1}^{n}(y_i - \hat{y}_i)^2 \tag{1}$$

$$= \sum_{i=1}^{n}\left(y_i - \beta_0 - \sum_{j=1}^{p}\beta_j x_{ij}\right) \tag{2}$$

- Ridge regression: minimize:

$$\sum_{i=1}^{n}\left(y_i - \beta_0 - \sum_{j=1}^{p}\beta_j x_{ij}\right) - \lambda\sum_{j=1}^{p}\beta_j^2 \tag{3}$$

$$= \text{RSS} - \lambda\sum_{j=1}^{p}\beta_j^2 \tag{4}$$

Ridge regression continued

- The second term, $\lambda \sum_{j=1}^{p} \beta_j^2$, is called a shrinkage penalty, and serves to shrink the estimates of β_j toward zero
  - when λ is large, β_j will shrink closer to zero, and when $\lambda \to \infty$, $\beta_j = 0$
  - when $\lambda = 0$, β_j is same as OLS solution
- Ridge regression will produce a different estimate of β_j for each value of λ
- So λ must be chosen carefully

The lasso

- Lasso: minimize:

$$\sum_{i=1}^{n}\left(y_i - \beta_0 - \sum_{j=1}^{p}\beta_j x_{ij}\right) - \lambda\sum_{j=1}^{p}|\beta_j| \tag{5}$$

$$= \text{RSS} - \lambda\sum_{j=1}^{p}|\beta_j| \tag{6}$$

- The lasso uses an $\ell 1$ penalty – the $\ell 1$ norm of a coefficient vector β is given by $\|\beta\|_1 = \sum|\beta_j|$
  (note: the ridge penalty is also known as the $\ell 2$ norm)

Differences

- Lasso can actually shrink some β values to zero completely, while ridge regression always includes them with some penalty
- This property makes interpreting the lasso simpler
- No steadfast rule as to which performs better in applications
  - depends on the number of predictors actually related to the outcome
  - depends on λ