Lec 14: Stepwise Regression

Ailin Zhang

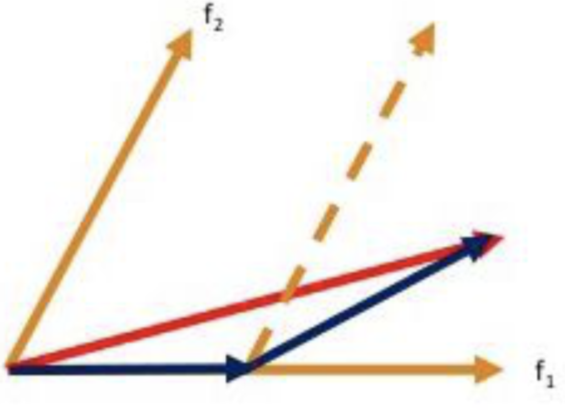
Roadmap for Regularized Learning

- Ridge regression
- Lasso regression
- Coordinate descent
- Spline regression
- Least angle regression
- Stagewise regression for ϵ learning
- Bayesian regression
- Perceptron
- SVM
- Adaboost

Least Angle Regression (LAR)

- LAR is a type of forward stepwise regression.
  - Forward step-wise regression builds a model sequentially, adding one variable at a time. At each step, it identifies the best variable to include in the active set, and then updates the least squares fit to include all the active variables.
- LAR is connected to Lasso regression.
- LAR was defined in the Efron et al., 2004. It is a relatively newer algorithm and is viewed as a democratic version of the forward step-wise regression.

Least Angle Regress (LAR)

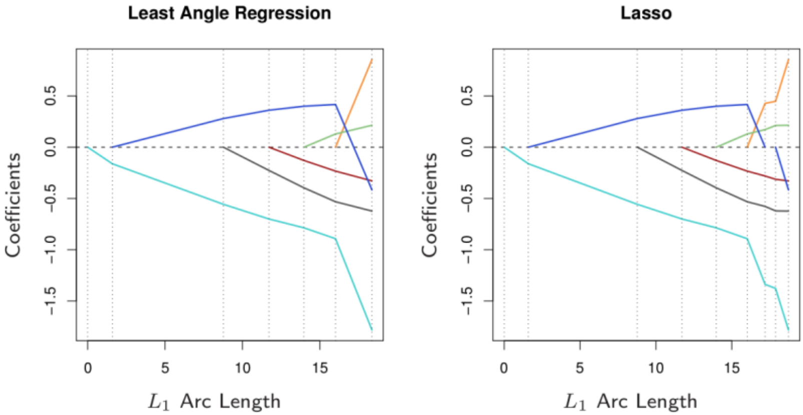


Least Angle Regression (LAR)

The basic steps of the Least-angle regression algorithm are:

1. Standardize the predictors (mean zero, unit norm). Initialize:
   - residuals $r = y - \bar{y}$
   - regression coefficient estimates $\beta_1 = \cdots = \beta_p = 0$;
2. find the predictor x_j most correlated with r;
3. move $\hat{\beta}_j$ towards its least-squares coefficient $\langle x_j, r \rangle$,
   - until for $k \neq j, \text{corr}\,(x_k, r) = \text{corr}\,(x_j, r)$.
4. add x_k in the active list and update both $\hat{\beta}_j$ and $\hat{\beta}_k$:
   - towards their joint least squares coefficient;
   - until x_l has as much correlation with the current residual;
5. continue until all p predictors have been entered.

LAR Solution Path



LAR vs Lasso

If β is the Lasso solution, at any given λ, let $\mathbf{R} = \mathbf{Y} - \sum_{j=1}^{p} \mathbf{X}_j \beta_j$, then $\hat{\beta}_j = \beta_j + \langle \mathbf{R}, \mathbf{X}_j \rangle / \|\mathbf{X}_j\|_{\ell_2}^2$. then

$$\langle \mathbf{R}, \mathbf{X}_j \rangle = \begin{cases} \lambda, & \text{if } \beta_j > 0, \\ -\lambda, & \text{if } \beta_j < 0, \\ s\lambda & \text{if } \beta_j = 0. \end{cases}$$

where $|s| < 1$.

Thus in the above process, for all of those selected $\mathbf{X}_j$, the algorithm maintains that $\langle \mathbf{R}, \mathbf{X}_j \rangle$ to be λ or $-\lambda$, for all selected $\mathbf{X}_j$. If we interpret $|\langle \mathbf{R}, \mathbf{X}_j \rangle|$ in terms of the angle between $\mathbf{R}$ and $\mathbf{X}_j$, then we may call the above process as least angle regression (LAR). In fact, the solution path is piecewise linear, and the LARS computes the linear pieces analytically instead of gradually reducing λ as in coordinate descent.

Lasso can be thought of as restricted versions of LAR

Other shrinkage methods: Group Lasso

Sometimes predictors belong to the same group: - genes that belong to the same molecular pathway; - dummy variables from the same categorical variable . . .

Suppose the p predictors are grouped in L groups, group lasso minimizes

$$\min_{\beta}\left\{\left\|y-\beta_0-\sum_{\ell=1}^{L}X_\ell\beta_\ell\right\|_2^2+\lambda\sum_{\ell=1}^{L}\sqrt{p_\ell}\|\beta_j\|_2\right\}.$$

where:

- $\sqrt{p_\ell}$ accounts for the group sizes;
- $\|\cdot\|$ denotes the (not squared) L2 norm
- sparsity is encouraged at both group and individual levels.

Other shrinkage methods: Non-negative garrote

The idea of lasso originates from the non-negative garrote,

$$\hat{\beta}_{\text{garrote}} = \text{argmin}_{\beta} \sum_{i=1}^{N} \left(y_i - \beta_0 - \sum_{j=1}^{p} c_j \beta_j x_{ij} \right)^2,$$

subject to

$$c_j \geqslant 0 \text{ and } \sum_j c_j \leqslant t$$

Non-negative garrote starts with OLS estimates and shrinks them:

- by non-negative factors;
- the sum of the non-negative factor is constrained;
- for more information, see Breiman (1995).

Stagewise Regression

The stagewise regression iterates the following steps:

1. Start with $\mathbf{R} = \mathbf{Y}$, $\beta_1, \beta_2, \cdots, \beta_p = 0$.
2. Find the predictor $\mathbf{X}_j$ most correlated with R: ind j with the maximal $|\langle \mathbf{R}, \mathbf{X}_j \rangle|$.
3. Then update $\beta_j \leftarrow \beta_j + \epsilon \langle \mathbf{R}, \mathbf{X}_j \rangle$
4. Set $\mathbf{R} = \mathbf{Y} - \sum_{j=1}^{p} \mathbf{X}_j \beta_j$, or $\mathbf{R} = \mathbf{R} - \epsilon \langle \mathbf{R}, \mathbf{X}_j \rangle \cdot \mathbf{X}_j$.

Repeat step 2-4

This is similar to the matching pursuit but is much less greedy. Such an update will change $\mathbf{R}$ and reduce $|\langle \mathbf{R}, \mathbf{X}_j \rangle|$, until another $\mathbf{X}_j$ catches up.

So overall, the algorithm ensures that all of the selected $\mathbf{X}_j$ to have the same $|\langle \mathbf{R}, \mathbf{X}_j \rangle|$.

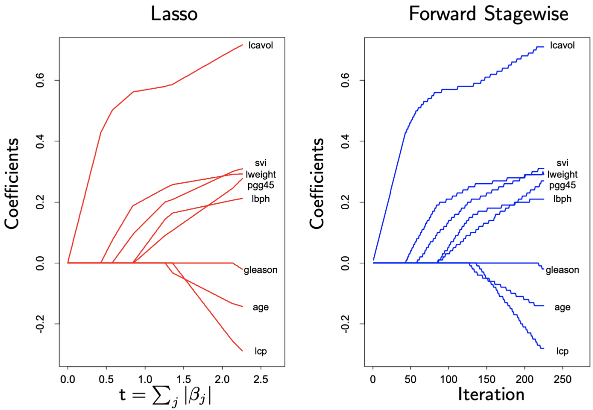
The stagewise regression is also called ϵ-boosting.

R code for Stagewise Regression

```
T = 3000
epsilon = .0001
beta = matrix(rep(0, p), nrow = p)
db = matrix(rep(0, p), nrow = p)
beta_all = matrix(rep(0, p*T), nrow = p)

R = Y
for (t in 1:T)
    {
      for (j in 1:p)
        db[j] = sum(R*X[, j])
      j = which.max(abs(db))
      beta[j] = beta[j]+db[j]*epsilon
      R = R - X[, j]*db[j]*epsilon
      beta_all[, t] = beta
}
matplot(t(matrix(rep(1, p), nrow = 1)%*%abs(beta_all)), t(beta_all), type = 'l')
```

Stagewise Regression vs Lasso Regression



Forward Stagewise and Lasso look similar. Are they Identical?

- If X is orthogonal: yes
- A more general case: almost identical, not exactly same.

Relationship among Lasso, LAR, and stagewise regression

- LAR: uses least squares directions in the active set of variables
- LASSO: uses least square directions; if a variable crosses zero, it is removed from the active set.
- Forward stagewise: Move in the direction of maximum $Corr(\mathbf{R}, \mathbf{X}_j)$ in the active set.

In forward stagewise, if $\epsilon \to 0$, it converges to LAR.

Stepwise Regression

- Stepwise regression is a **variable selection** procedure for independent variables (**X**)
- Consists of a series of steps designed to find the most features to include in a regression model
- Basis for selection:
    - Choose a variable that satisfies the criterion
    - Remove a variable that least satisfies the criterion

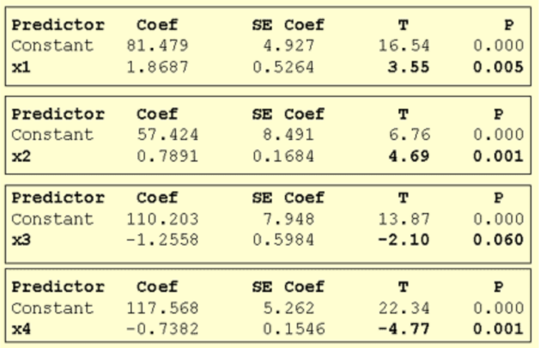
Stepwise Regression: Example

At each step, we either enter or remove a predictor based on the partial F-tests — the t-tests for the slope parameters.

We stop when no more predictors can be justifiably entered or removed from our stepwise model, thereby leading us to a “final model.”

Stepwise Regression: Example

Regress y on x_1, y on x_2, y on x_3, y on x_4. Choose significance level as 0.15.



```
Predictor   Coef       SE Coef      T        P
Constant    81.479      4.927      16.54    0.000
x1          1.8687     0.5264       3.55    0.005
```

```
Predictor   Coef       SE Coef      T        P
Constant     57.424     8.491       6.76    0.000
x2           0.7891    0.1684       4.69    0.001
```

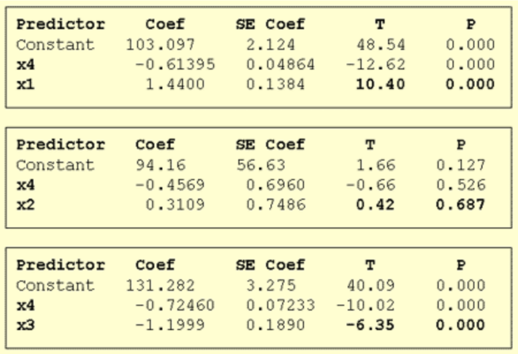
```
Predictor   Coef       SE Coef      T        P
Constant    110.203     7.948      13.87    0.000
x3          -1.2558    0.5984      -2.10    0.060
```

```
Predictor   Coef       SE Coef      T        P
Constant    117.568     5.262      22.34    0.000
x4          -0.7382     0.1546     -4.77    0.001
```

As a result of the first step, we enter x_4 into our stepwise model.

Stepwise Regression: Example

we fit the next two-predictor model that includes x_4 as a predictor — that is, we regress y on x_4 and x_1, y on x_4 and x_2 , and y on x_4 and x_3



```
Predictor     Coef      SE Coef       T        P
Constant    103.097     2.124       48.54    0.000
x4           -0.61395   0.04864    -12.62    0.000
x1            1.4400    0.1384      10.40    0.000
```

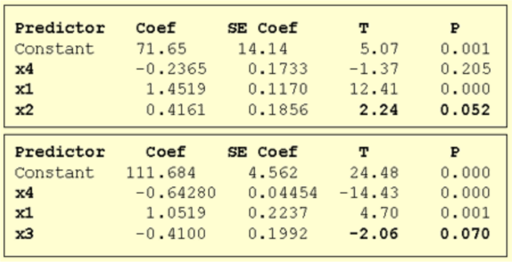
```
Predictor    Coef       SE Coef       T       P
Constant     94.16      56.63        1.66   0.127
x4           -0.4569    0.6960      -0.66   0.526
x2            0.3109    0.7486       0.42   0.687
```

```
Predictor    Coef       SE Coef       T       P
Constant    131.282     3.275       40.09   0.000
x4           -0.72460   0.07233    -10.02   0.000
x3           -1.1999    0.1890      -6.35   0.000
```

As a result of the second step, we enter x_1 into our stepwise model.

Stepwise Regression: Example

Regress y on x_4, x_1,and x_2, and y on x_4, x_1 and x_3



```
Predictor    Coef      SE Coef       T        P
Constant     71.65      14.14       5.07    0.001
x4          -0.2365     0.1733     -1.37    0.205
x1           1.4519     0.1170     12.41    0.000
x2           0.4161     0.1856      2.24    0.052
```

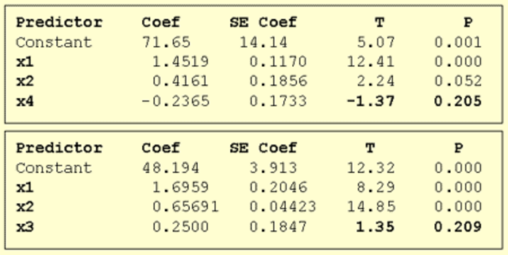
```
Predictor     Coef     SE Coef       T        P
Constant    111.684     4.562      24.48    0.000
x4          -0.64280    0.04454   -14.43    0.000
x1           1.0519     0.2237      4.70    0.001
x3          -0.4100     0.1992     -2.06    0.070
```

As a result of the third step, we enter x_1 into our stepwise.

At the same time, remove x_4.

Stepwise Regression: Example

Proceed fitting each of the three-predictor models that include x_1 and x_2 as predictors — that is, we regress y on x_1, x_2, and x_3; y on x_1, x_2, and x_4:



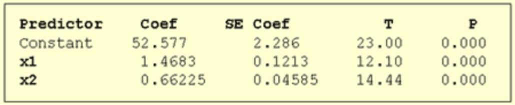
Predictor	Coef	SE Coef	T	P
Constant	71.65	14.14	5.07	0.001
x1	1.4519	0.1170	12.41	0.000
x2	0.4161	0.1856	2.24	0.052
x4	-0.2365	0.1733	**-1.37**	**0.205**

Predictor	Coef	SE Coef	T	P
Constant	48.194	3.913	12.32	0.000
x1	1.6959	0.2046	8.29	0.000
x2	0.65691	0.04423	14.85	0.000
x3	0.2500	0.1847	**1.35**	**0.209**

We stop our stepwise regression procedure. Our final regression model, based on the stepwise procedure contains only the predictors x_1 and x_2.

Stepwise Regression: Example

Final model:



```
Predictor     Coef      SE Coef        T        P
Constant    52.577       2.286     23.00    0.000
x1           1.4683      0.1213    12.10    0.000
x2           0.66225     0.04585   14.44    0.000
```

Not only can you use t-test, you can also consider R^2, AIC, BIC,etc...

Summary for Stepwise Regression

1. The final model is not guaranteed to be optimal in any specified sense.
2. The procedure yields a single final model, although there are often several equally good models.
3. Stepwise regression does not take into account domain knowledge about the predictors. It may be necessary to force the procedure to include important predictors.
4. One should not over-interpret the order in which predictors are entered into the model.
5. It is possible that we may have committed a Type I or Type II error along the way.