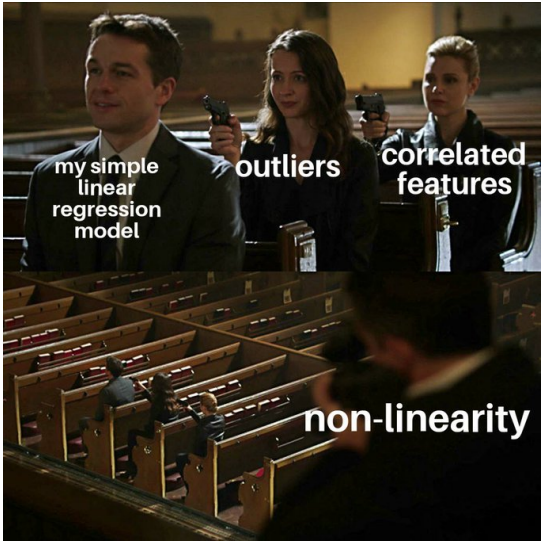
BIOS 825: LR Model Selection

Alexander McLain

September 5, 2023

Outline

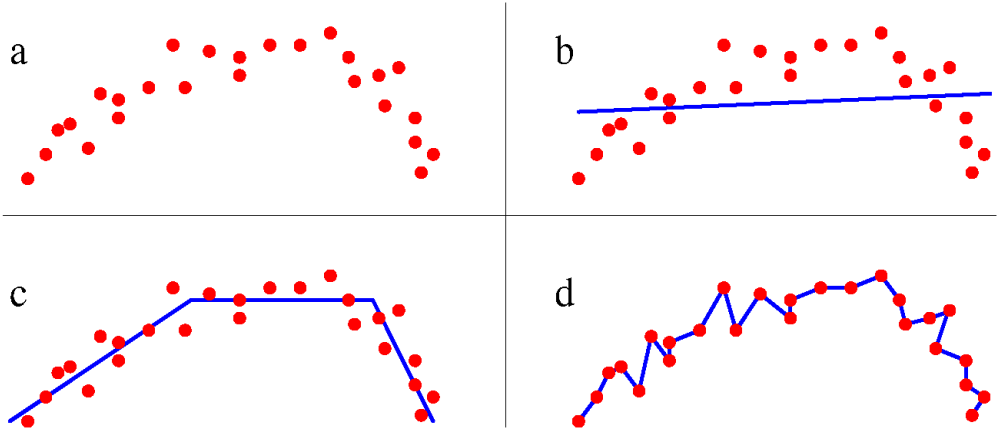

my simple linear regression model
outliers
correlated features
non-linearity

Why Model Selection?

1. *Prediction accuracy:* to make reasonable predictions or estimations, we need
   - **accuracy** $\rightarrow$ on average, what we estimate is equal to what we expect (e.g., $\hat{Y}$ in the long run is equal to the population mean of Y)
   - **precision** $\rightarrow$ small variation in prediction/estimation
2. *Interpretation:* if we can limit the number of variables, we can better understand the "main factors" driving the outcome.

For this section, we will assume that the assumptions of the model are met and focus on finding the best model.

Underfitting vs. Overfitting



Underfitting vs. Overfitting

- Underfitting occurs when important regressors are left out of the model. Costs:
  - deficient models (i.e., missing patterns).
  - misinterpretations of variable relationships.
- Overfitting occurs when all-important regressors are in the model, but some unimportant ones are, too. Costs:
  - df for error
  - unneeded complexity and increased variance of the predicted values.
  - somewhat widened confidence and prediction intervals.

Best-Subset Selection

- Best-subset can look over all 2^p models or look at it for only those with k variables.
- For the latter, this method finds for each k the subset of k variables that minimizes the RSS.
- Even for p as large as 30 or 40 there are algorithms to estimate this quickly.
- This procedure cannot be used to estimate k (RSS will necessarily go down with k)
- To choose k we need to use some other criteria.

Model Selection Criteria

1. MSE
   - If the model underfits the data $\rightarrow$ large residuals $\rightarrow$ large SSE $\rightarrow$ large MSE.
   - If the model overfits the data $\rightarrow$ df for the error part is lost $\rightarrow$ large MSE
   - A good model is expected to have a relatively small MSE.
2. Adjusted R^2 (R^2_{Adj} in SAS, or $\bar{R}^2$)
   - $\bar{R}^2 = 1 - \frac{MSE}{SS_{TOTAL}/(n-1)}$
   - This is similar to R^2, but we are "punished" (via df) for using too many terms in the model
   - Since SS_{TOTAL} is not changed between models, as MSE decreases, $\bar{R}^2$ increases.
   - In SAS, use "selection=adjrsq" in PROC REG if want to use this criterion for model selection

$*$ The two criteria MSE and $\bar{R}^2$ are equivalent.

Model Selection Criteria

3. Mallow's C(p) or C_p
   - This criterion looks at both accuracy and precision.
   - p: number of terms in the model under consideration, $p \leq k$ (k: the number of independent variables in the full model).
   - $\hat{\sigma}^2$: MSE from the full model (with k independent variables).
   - $MSE(p)$: MSE from the model under consideration.
   - Def. of $C(p)$

$$\begin{aligned} C(p) &= (p+1) + \frac{(MSE(p) - \hat{\sigma}^2)(n-p-1)}{\hat{\sigma}^2} \\ &= SSE(p)/\hat{\sigma}^2 - [n - 2(p+1)] \end{aligned}$$

   - a good model has C(p) close to (p+1)
   - Mallow's Cp is equivalent to Akaike information criterion (AIC)
     $\mathrm{AIC} = 2(p+1) - 2\log(\hat{L})$ for normal errors where $\hat{L}$ is the fitted normal likelihood.

Model Selection Criteria

4. PRESS (prediction SS)
   - This criterion focuses on quality of prediction for a new data point not included in the current data set.
   - The idea is to remove one point from the data, use the remaining data to fit the model, and then predict the removed point.
   - $\hat{Y}_{(i)}$: the predicted value for the i-th observation (data point i is not used for estimating the coefficients).
   - $PRESS = \sum_{i=1}^{n}(Y_i - \hat{Y}_{(i)})^2$
   - A good model has a small PRESS statistic.
   - In SAS, use the PRESS option in the OUTPUT statement in PROC REG to get $Y_i - \hat{Y}_{(i)}$ for all subjects.
5. Cross validation of two data sets $\rightarrow$ the concept is similar to PRESS.

Automated Model Selection Procedures

- Useful when there are a large number of possible independent variables.
- Can use these procedures to get a manageable number of "candidate models".
- These procedures start with a full or null model and automatically add or subtract variables.
- Computationally, these methods are cheap.
- We don't look at as many models (more constrained search): results will have lower variance, but possibly more bias.
- They depend on criteria to evaluate models, traditional software will use p-values (which is not recommended) here we'll consider AIC.

Automated Model Selection Procedures

1. Forward selection
   a. begin with nothing (except for the intercept) in the model
   b. enter the variable with the highest R^2 (SLR) (assuming it increases AIC over a null model)
   c. determine which variable would result in the highest increase in AIC and add it to the active set of variables
   e. repeat step c until no more variables increase the AIC
2. Backward selection
   a. begin with all variables in the model
   b. Delete the variable that would result in the highest increase in AIC
   c. repeat step b until no variable would improve AIC

Automated Model Selection Procedures

3. Stepwise selection
   a. Enter the variable with the highest R^2 (SLR) (assuming it increases AIC over a null model)
   b. At each step, check which variable would result in the highest increase in AIC and add it to the active set of variables
   c. check if removing any variable would improve the AIC, if there is such a variable remove it from the active set
   d. repeat steps b and c until no variable improves AIC or a loop is encountered.

Forward-Stagewise Regression

- The Forward-Stagewise Regression algorithm can be considered as a "slow" version of the forward selection algorithm.[1]
- As we'll see, this algorithm ends up being closely related to the LASSO and LAR methods.
- Thus, a good understanding of this method is important as we move forward.
- For this algorithm (and many others) we'll assume that the $\boldsymbol{X}$ vector has been standarized.

[1] The description of Forward-Stagewise in ESL is poor. See Efron, Hastie, Johnstone and Tibshirani (2003) "Least Angle Regression" (with discussion) *Annals of Statistics* for a better treatment of the method.

The Forward-Stagewise Regression algorithm

1. Let $\hat{\boldsymbol{y}}^{(0)} = \bar{y}$, $\hat{\boldsymbol{\beta}}^{(0)} = \mathbf{0}$, and $\boldsymbol{r}^{(0)} = \boldsymbol{y} - \hat{\boldsymbol{y}}^{(0)}$
2. For step $k = 1, 2, \ldots$
   - 2.1 let $\boldsymbol{c}^{(k)} = \boldsymbol{X}'(\boldsymbol{y} - \hat{\boldsymbol{y}}^{(k-1)})$ be the vector of *current correlations*
   - 2.2 find $j^* = \text{argmax}_j |c_j^{(k)}|$ be the $\boldsymbol{X}$ variable with the largest current correlation.
   - 2.3 Let $\delta = \epsilon \cdot \text{sign}(c_{j^*}^{(k)})$and set

$$\begin{aligned} \hat{\beta}_{j^*}^{(k)} &= \hat{\beta}_{j^*}^{(k-1)} + \delta \\ \hat{\boldsymbol{y}}^{(k)} &= \hat{\boldsymbol{y}}^{(k-1)} + \delta \boldsymbol{X}_{j^*} \end{aligned}$$

   where ϵ is a "small" constant.

repeat step 2 until $\max_j |c_j^{(k)}| \leq \kappa$

The Forward-Stagewise Regression algorithm notes

- If $\kappa = 0$ the algorithm will continue until we reach the LS estimates
- The choice of ϵ (the step-size) is important.
- If $\epsilon = \max_j |c_j^{(k)}|$ leads to classic forward selection (if *RSS* is used).
- This procedure can take many more than p steps.
- The basic idea is to combine many "weak" models (i.e., ϵ is small) to build a powerful "committee" model
- That is, we find the variable that best describes the data to the residuals from our current model, our current set of errors
- **Boosting** (a common machine learning approach) is also referred to as a "slow learning" algorithm.