Introduction; problems with classical methods

Patrick Breheny

January 14

Introduction

- This course concerns the analysis of data in which we are attempting to predict an outcome Y using a number of explanatory factors $X_1, X_2, X_3, \ldots$, some of which may not be particularly useful
- Although the methods we will discuss can be used solely for prediction (i.e., as a "black box"), I will adopt the perspective that we would like the statistical methods to be interpretable and to explain something about the relationship between the X and Y
- Regression models are an attractive framework for approaching problems of this type, and the majority of the course will focus on extending classical regression modeling to deal with high-dimensional data

High-dimensional data

Modern computation has changed the way science is conducted, and enabled researchers to easily collect, store, and access data for large numbers of features (ballpark number of features in parentheses):

- Advances in information technology such as REDCap (~ 100)
- Adoption of electronic medical records (> 100)
- Molecular biology technologies such as microarrays and RNA-Seq ($> 10,000$)
- Advances in genotyping and genetic sequencing ($> 100,000$)

High-dimensional data (cont'd)

- This type of data is known as *high dimensional data*
- Throughout the course, we will let
  - n denote the number of independent sampling units (e.g., number of patients)
  - p denote the number of features recorded for each unit
- In high-dimensional data, p is large with respect to n
  - This certainly includes the case where $p > n$
  - However, the ideas we discuss in this course are also relevant to many situations in which $p < n$; for example, if $n = 100$ and $p = 80$, we probably don't want to use ordinary least squares

More notation

- We will use $\mathbf{X}$ denote the $n \times p$ matrix containing the predictor variables, with element x_{ij} recording the value of the jth feature for the ith independent unit
- We will let $\mathbf{y}$ denote the length-n vector of response values
- For the sake of simplicity, for most of the course we will assume that Y is normally distributed, but we will consider other types of responses in the "Other likelihoods" topic

Univariate analysis

- A simple, widely used approach to analyzing high-dimensional data is to split the problem up into a large number of low-dimensional problems
- For example, rather than trying to regress $\mathbf{y}$ simultaneously on all the features, we can carry out p separate single-variable regressions, one for each feature:

$$y_i = \alpha_j + \beta_j x_{ij} + \epsilon_i$$
$$\epsilon_i \overset{\perp\!\!\!\perp}{\sim} \mathrm{N}(0, \sigma^2);$$

this approach is also known as *marginal regression*

Univariate analysis: Challenges

- The appeal of this approach is classical regression can be easily applied to the separate analyses to yield estimates $\{\widehat{\beta}_j\}$, confidence intervals, and test hypotheses to produce p-values $\{p_j\}$
- The major complication, however, is that this approach involves a large number of separate analyses that must somehow be combined into a single set of results
- Thus, while standard methods can be used for the initial analyses, there has been a great deal of innovation over the past 30 years in terms of how to combine these results; we will discuss these innovations during the "Large scale testing" topic

Limitations of univariate models

Marginal regression is straightforward, but has several drawbacks:

- Fails to account for correlation among the features
- Provides no way to estimate the independent effect of a feature while other features remain unchanged
- Diminished power
- No good way to combine the predictions of separate regressions into a single overall prediction
- No way of assessing the overall proportion of the variability in the outcome that may be explained by the features

Joint modeling

- These issues can only be resolved by considering a joint model of the relationship between $\mathbf{y}$ and the full set of features:

$$y_i = \beta_0 + \sum_{j=1}^{p} \beta_j x_{ij} + \epsilon_i$$
$$\epsilon_i \overset{\perp\!\!\!\perp}{\sim} \mathrm{N}(0, \sigma^2)$$

- The maximum likelihood approach involves solving for the value of $\boldsymbol{\beta}$, known as the maximum likelihood estimator (MLE), that minimizes the residual sum of squares $\|\mathbf{y} - \mathbf{X}\boldsymbol{\beta}\|^2$
- Here, $\|\mathbf{v}\| = \sqrt{\sum_i v_i^2}$ denotes the Euclidean norm; we will use this notation frequently throughout the course

OLS

- The solution is determined by the linear system of equations

$$\mathbf{X}^T\mathbf{X}\widehat{\boldsymbol{\beta}} = \mathbf{X}^T\mathbf{y}$$

- Provided that $\mathbf{X}^T\mathbf{X}$ is invertible, the system has the unique solution

$$\widehat{\boldsymbol{\beta}} = (\mathbf{X}^T\mathbf{X})^{-1}\mathbf{X}^T\mathbf{y},$$

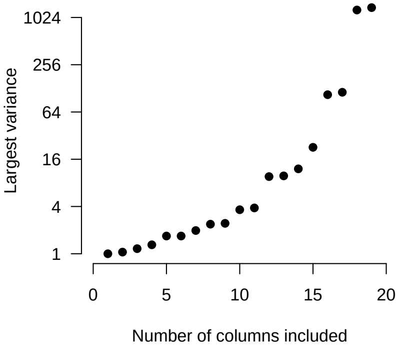
  known as the *ordinary least squares* (OLS) estimate
- The OLS estimate resolves all of the issues on slide 8 and has many well-recognized benefits such as yielding best linear unbiased estimates of $\boldsymbol{\beta}$

MLE problems

- However, there are many drawbacks to the use of maximum likelihood for estimating $\boldsymbol{\beta}$ when p is large
- Most dramatically, when $p \geq n$ the matrix $\mathbf{X}^T\mathbf{X}$ is not invertible and the MLE is not unique
- However, even if $\mathbf{X}^T\mathbf{X}$ can be inverted and a unique maximum identified, as p increases and $\mathbf{X}^T\mathbf{X}$ approaches singularity, the likelihood surface becomes very flat
- This means that a wide range of values of $\boldsymbol{\beta}$ are consistent with the data and wide confidence intervals required to achieve, say, 95% coverage

An example

Consider a matrix $\mathbf{X}$ with $n = 20$ and whose elements consist of independent, normally distributed random numbers; the figure below plots the largest variance of the $\widehat{\beta}_j$ estimates as we increase the number of columns in $\mathbf{X}$:



Remarks

- As $p \to n$, $\mathbb{V}(\widehat{\boldsymbol{\beta}})$ increases without bound; the increase is substantial as p approaches n, and infinite when $p \geq n$
- Clearly, maximum likelihood cannot accommodate high-dimensional data without running into serious problems of identifiability and inefficiency

The oracle model

- Suppose, however, that many features are unrelated to the outcome (in the sense that $\beta_j = 0$), and only a few features are important
- If we knew in advance which elements of $\boldsymbol{\beta}$ are zero and which are not, then we could modify maximum likelihood without abandoning it completely, and avoid all of the earlier problems
- Specifically, we could apply maximum likelihood only to the variables for which $\beta_j \neq 0$; this is known as the *oracle* model

Model selection

- Obviously, the oracle model is a theoretical gold standard, not a realistic approach to data analysis, as it would require access to an oracle that could tell you which features are related to the outcome and which are not
- In the real world, we have to use the data in order to make empirical decisions about which features are related to the outcome and which are not; this is known as *model selection*

The model selection problem

- Unfortunately, using the same data for two purposes – to select the model and also to carry out inference with respect to the model's parameters – introduces substantial biases and invalidates the inferential properties that maximum likelihood typically possesses
- To illustrate, consider the following simulation:

$$x_{ij} \overset{\perp\!\!\!\perp}{\sim} \text{Unif}(0,1) \qquad \text{for } j \text{ in } 1, 2, \ldots, 100$$
$$y_i \overset{\perp\!\!\!\perp}{\sim} \text{N}(0,1)$$

for i in $1, 2, \ldots, 25$

- We will use BIC to select the 5 most important variables, then use OLS with only those variables, and repeat this 100 times

MSE, MSPE, and ME

- Before we show the results, let's define three quantities that we will use throughout the semester
- The first is the mean squared (estimation) error, or MSE:

$$\text{MSE} = \mathbb{E}\|\widehat{\boldsymbol{\beta}} - \boldsymbol{\beta}\|^2$$

  Note: "mean" here refers to the expected value, not to averaging over the number of terms we are estimating
- The second is the mean squared prediction error, or MSPE:

$$\text{MSPE} = \frac{1}{n}\sum_{i=1}^{n}(y_i - f(\mathbf{x}_i))^2 \qquad \text{(sample)}$$

$$\text{MSPE} = \mathbb{E}\{(Y - f(\mathbf{x}))^2|\mathbf{x}\} \qquad \text{(population)}$$

  Note: For linear regression, $f(\mathbf{x}_i) = \mathbf{x}_i^T\widehat{\boldsymbol{\beta}}$; from context, it is usually clear which one we are talking about, but if necessary, I will denote the former $\widehat{\text{MSPE}}$

MSE, MSPE, and ME (cont'd)

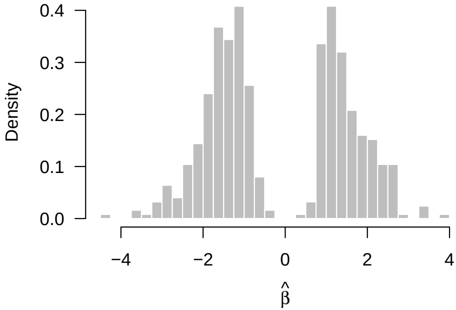
- Lastly, we will sometimes refer to the model error (ME):

$$\text{ME} = \{\mathbb{E}(y|\mathbf{x}) - f(\mathbf{x})\}^2$$

- Note that for (homoskedastic) linear regression models, $\text{MSPE} = \text{ME} + \sigma^2$, but the distinction is useful when considering consistency, since ideally ME will go to zero for a model as we collect more data, but it is impossible for MSPE to go to zero
- Finally, note that in random $\mathbf{x}$ settings, all three quantities would have an additional outer expectation over $\mathbf{x}$ (keep this in mind for your simulations)

Results

A histogram of the 500 $\widehat{\beta}_j$ estimates we obtain:



Remarks

- As we will see, this approach performs terribly
- By using the data set for model selection as well as estimation and inference, we have grossly distorted the sampling distribution of $\widehat{\beta}$
- This has dramatic consequences in terms of estimation, prediction, variable selection, and the validity of inference

Estimation

- The model selection process heavily biases the estimates of the regression coefficients away from zero
- In our simulation, most estimates were approximately ± 1.5 instead of being close to 0, the true value
- In particular, the average MSE is 2.7, compared to 0.48 for marginal regression, roughly a 5-fold increase
- This phenomenon is sometimes referred to as the "winners' curse"

Variable selection

- Here, we imposed an upper bound of 5 on the number of variables we allowed to be selected by the BIC-guided forward selection process; in all 100 replications, this upper bound was reached
- Obviously, since the true model in this case is the null (intercept-only) model, the model selection process we have employed here results in systematic overfitting
- While it is true that asymptotically, BIC will select the true model with probability tending to 1, that asymptotic argument relies on p remaining fixed while $n \to \infty$, or in other words, on $n \gg p$
- Clearly, BIC cannot be relied on for accurate variable selection in high-dimensional problems

Prediction

- On average, the selected models achieved a mean squared prediction error of 2.15, compared to a prediction error of $\sigma^2 = 1$ for the null model
- Thus, by carrying out model selection, we have reduced the predictive accuracy of the model by half (doubled its error)

Inference

- Finally, let us consider the validity of the inferences that we obtain from the post-selection OLS model:
  - The median p-value for testing $H_0 : \beta_j = 0$ was $p = 0.0013$
  - The actual coverage achieved by constructing 95% confidence intervals was under 5%
- Ignoring selection effects when carrying out post-selection inference produces conclusions that are far too liberal, with actual errors accumulating at a much higher rate than the statistical inferential approaches would indicate
- In summary, this approach is wildly optimistic and overconfident

Final remarks

- These problems are widely recognized; unfortunately, they are also widely ignored
- The problem of developing statistical methods capable of simultaneous variable selection and inference has challenged statisticians for decades, from Scheffé (1953) to the present
- One of the primary goals of this course is to demonstrate the extent to which recent developments in penalized regression address and alleviate the concerns about simultaneous selection and inference we have raised today