6.4 Considerations in High Dimensions

High-Dimensional Data

- Traditional statistical techniques are typically for low-dimensional data, where $n >> p$
  - Historically, most data looked like this.
- Modern data is often far more complex.
  - Technology has changed the way data is collected.
  - We have more access to more information than ever before.
  - Often, we can collect huge numbers of features p, but n is limited for practical reasons (cost, etc.)

Example

Rather than predicting blood pressure on the basis of just age, sex, and BMI, one might also collect measurements for half a million single nucleotide polymorphisms (SNPs; these are individual DNA mutations that are relatively common in the population) for inclusion in the predictive model. Then $n \approx 200$ and $p \approx 500,000$.

Example

- A marketing analyst interested in understanding people's online shopping patterns could treat as features all of the search terms entered by users of a search engine. This is sometimes known as the "bag-of-words" model.
- The same researcher might have access to the search histories of only a few hundred or a few thousand search engine users who have consented to share their information with the researcher.
- For a given user, each of the p search terms is scored present (0) or absent (1), creating a large binary feature vector. Then $n \approx 1,000$ and p is much larger.

High-Dimensional Data

- Data with more features than observtaions is referred to as *high-dimensional*.
- Classical approaches like least-squares regression will not work in this setting.
- Some issues include the bias-variance tradeoff and the danger of overfitting.
    - These are even more important with high-dimensional data.
- We will discuss considerations for $n < p$ and $n \approx p$.
    - Also worth considering when $n > p$

What Goes Wrong?

- When $n \leq p$, least squares will yield coefficient estimates that result in a perfect fit to the data.
  - Obviously an overfitting problem $\rightarrow$ causes least squares to be *too flexible*.
  - Messes up our ability to calculate standard errors, p-values, etc.
- If we don't examine possible overfitting, we may not notice that our model doesn't work.

Example 6.22

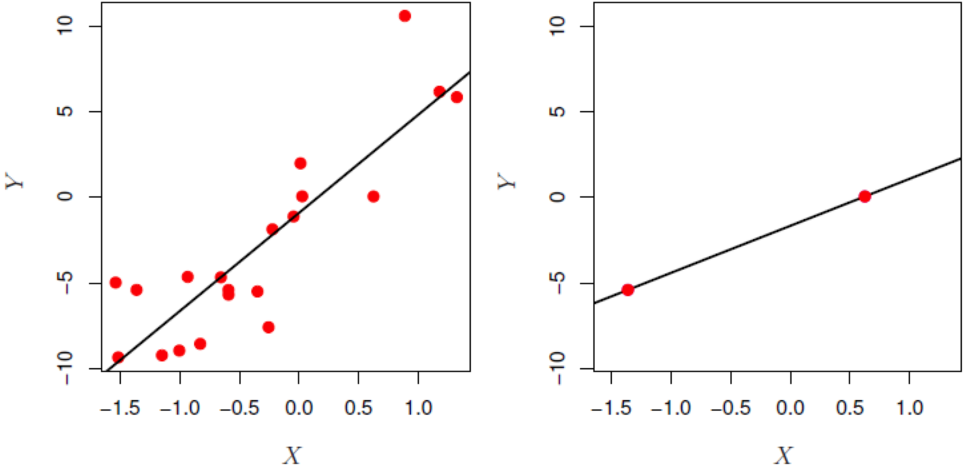


FIGURE 6.22. Left: *Least squares regression in the low-dimensional setting.* Right: *Least squares regression with $n = 2$ observations and two parameters to be estimated (an intercept and a coefficient).*

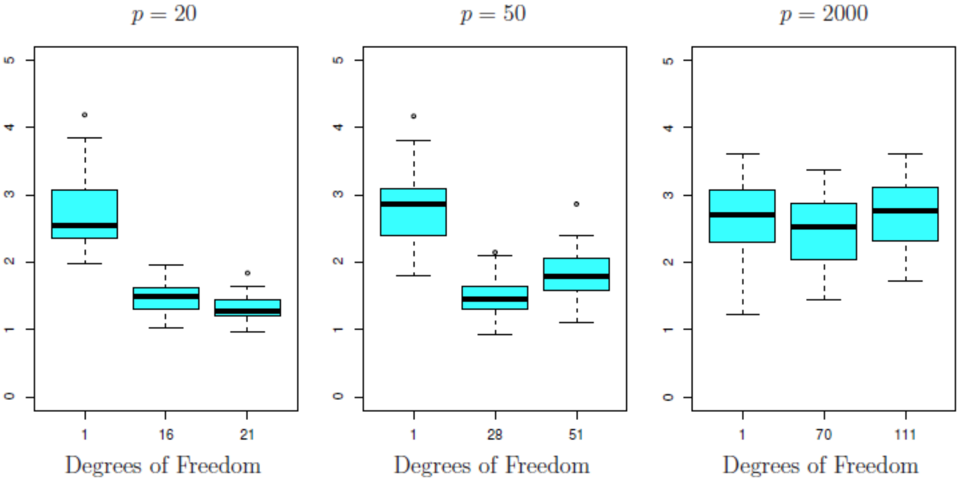
Another Challenge

- In the high-dimensional setting, AIC, BIC and R^2_{adj} don't work well.
  - The formula for $\hat{\sigma}^2$ yields an estimate of 0
  - Can obtain models with $R^2_{adj} = 1$
- So we need other approaches.
- Cross validation should still be good!

Regression in High Dimensions

- The methods seen in this chapter for fitting *less flexible* least squares models are very useful in this setting.
  - Help avoid overfitting by using a less flexible fitting approach.

Example 6.24



$p = 20$
$p = 50$
$p = 2000$
0
1
2
3
4
5
1
16
21
1
28
51
1
70
111
Degrees of Freedom
Degrees of Freedom
Degrees of Freedom

Example 6.24 Figure Details

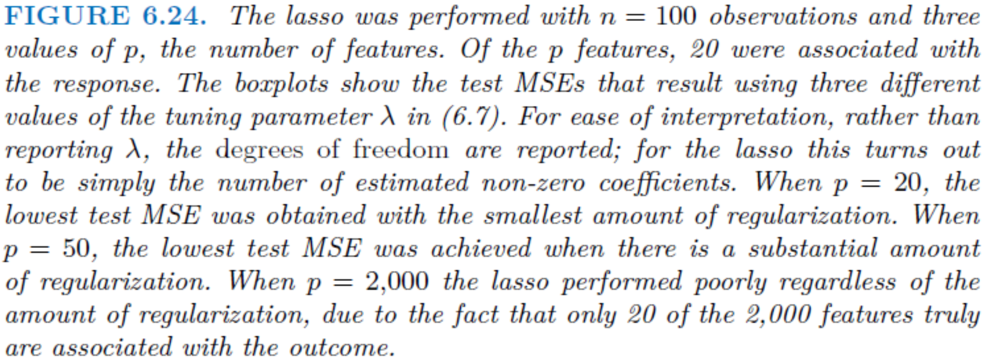


FIGURE 6.24. *The lasso was performed with $n = 100$ observations and three values of p, the number of features. Of the p features, 20 were associated with the response. The boxplots show the test MSEs that result using three different values of the tuning parameter λ in (6.7). For ease of interpretation, rather than reporting λ,* the degrees of freedom *are reported; for the lasso this turns out to be simply the number of estimated non-zero coefficients. When $p = 20$, the lowest test MSE was obtained with the smallest amount of regularization. When $p = 50$, the lowest test MSE was achieved when there is a substantial amount of regularization. When $p = 2{,}000$ the lasso performed poorly regardless of the amount of regularization, due to the fact that only 20 of the 2,000 features truly are associated with the outcome.*

Example 6.24

This highlights three important points:

1. Regularization or shrinkage plays a key role in high-dimensional problems.
2. Appropriate tuning parameter selection is crucial for good predictive performance.
3. The test error tends to increase as the dimensionality of the problem increases (unless the additional features are truly associated with the response)

The Curse of Dimensionality

- The temptation is to assume that more features means more information means better fit.
- But this may not be the case.
- In general, adding additional features will improve the model only if they are truly associated with the response.
- Adding features just to throw stuff at the wall and see what sticks can lead to a deterioration of the fitting model.
    - adds noise without adding useful information
    - increased overfit chance and so test set error
- Collecting tons of data is really only useful if the measurements have predictive relevance (and they often are not collected so thoughtfully)

Interpreting Results in High Dimensions

- We need to be cautious interpreting results for regression procedures in a high-dim setting
  - (lasso, ridge, etc)
- Multicollinearity can be extreme in the high-dim setting!
  - If we have enough variables, any variable can almost certainly be written as a linear combination of all of the other variables in the model.
  - So we can never know which variables (if any) are *truly* predictive of the outcome.
  - and we can never identify the *best* coefficient estimates
- This makes interpretation potentially problematic.

Example

- Suppose we are trying to predict blood pressure on the basis of half a million SNPs.
- Forward stepwise selection indicates that 17 of those SNPs lead to a good predictive model on the training data.
- We cannot conclude that these 17 SNPs predict blood pressure more effectively than the other SNPs not included in the model.
  - There are likely to be many sets of 17 SNPs that would predict blood pressure just as well as the selected model.

Example

- If we were to obtain a new data set and perform forward stepwise selection on it, we would likely obtain a model containing a different, and perhaps even non-overlapping, set of SNPs.
- This doesn't detract from the value of the model obtained.
    - The model might turn out to be very effective in predicting blood pressure on an independent set of patients, and might be clinically useful for physicians.
    - But we must be careful not to overstate the results.

Interpreting Results in High Dimensions

- Also important to be careful in reporting errors and measures of model fit.
- When p $>$ n, it is easy to obtain a useless model that has zero residuals.
  - So we should never use sum of squared errors, p-values, R2 statistics, or other traditional measures of model fit on the training data as evidence of a good model fit in the high-dimensional setting.
  - Ex: we may obtain a model with $R^2 = 1$ when $p > n$
    - Reporting this suggests that a statistically useful model has been obtained, whereas in fact this provides absolutely no evidence of a compelling model.
- Instead, we should report results on an independent test set, or cross-validation errors.