Lecture Notes

8. Collinearity and Model Selection

1. Introduction

- When there is a perfect linear relationship among the regressors in a linear model, the least-squares coefficients are not uniquely defined.
- A strong, but less than perfect, linear relationship among the X's causes the least-squares coefficients to be unstable:
  - Coefficient standard errors are large, reflecting the imprecision of estimation of the β's;
  - consequently confidence intervals for the β's are broad.
  - See Figure 1.

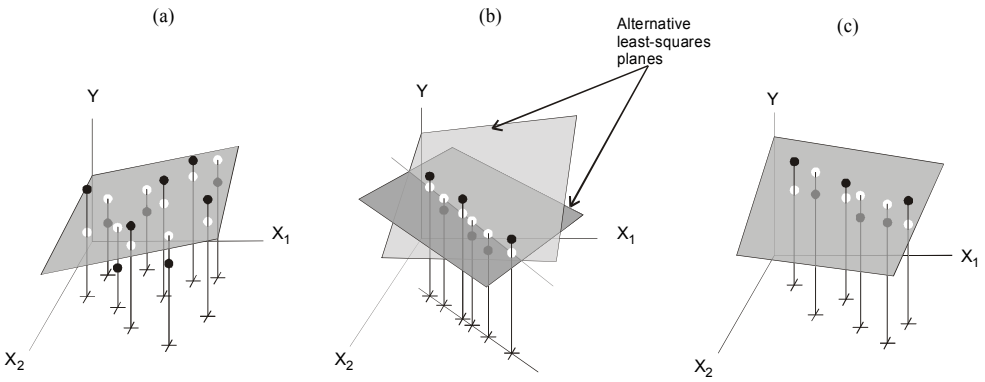

Figure 1. (a) Low correlation between X_1 and X_2 – regression plane well supported; (b) perfect correlation between X_1 and X_2, showing two of the infinite number of least-squares planes; (c) high but not perfect correlation between X_1 and X_2 – regression plane not well supported.

- Caveats:
  - Except in certain contexts — such as timeseries regression — collinearity is not a common problem in social-science applications of linear models.
    - Insufficient variation in explanatory variables, small samples, and large error variance are more frequently the source of imprecision in estimation.
  - Methods that are commonly employed as cures for collinearity — in particular, biased estimation and variable selection — can easily be worse than the disease.

- The detection of collinearity may not have practical implications.
  - The standard errors of the regression estimates are the bottom line:
  - If these estimates are precise, then the degree of collinearity is irrelevant.
  - If the estimates are imprecise, then knowing that the culprit is collinearity is of use only if the study can be re-designed to decrease the correlations among the X's, which is usually impossible in observational research.

2. Goals:

- To explain the nature of the collinearity 'problem' in regression.
- To introduce simple diagnostics for measuring collinearity.
- To describe several 'solutions' to the collinearity problem and to gain an appreciation of their limitations.
- To consider criteria for selecting statistical models in a more general framework.

3. Detecting Collinearity

- To summarize what we know:
  - When there is a perfect linear relationship among the X's,
$$c_1X_{i1} + c_2X_{i2} + \cdots + c_kX_{ik} = c_0$$
where $c_1, c_2, \ldots, c_k$ are not all 0:
    - The least-squares normal equations do not have a unique solution.
    - The sampling variances of the regression coefficients are infinite.
  - Perfect collinearity is usually the product of some error in formulating the linear model, such as failing to employ a baseline category in dummy regression.

- When collinearity is less than perfect:
  - The sampling variance of the least-squares slope coefficient B_j is
$$V(B_j) = \frac{1}{1 - R_j^2} \times \frac{\sigma_\varepsilon^2}{(n-1)S_j^2}$$
where
    - R_j^2 is the squared multiple correlation for the regression of X_j on the other X's,
    - and $S_j^2 = \sum(X_{ij} - \overline{X}_j)^2/(n-1)$ is the variance of X_j.
  - The term $1/(1 - R_j^2)$, called the *variance-inflation factor (VIF)* indicates the impact of collinearity on the precision of B_j.
  - The width of the confidence interval for β_j is proportional to the square root of the VIF.
  - It is not until R_j approaches .9 that the precision of estimation is halved (see Figure 2).

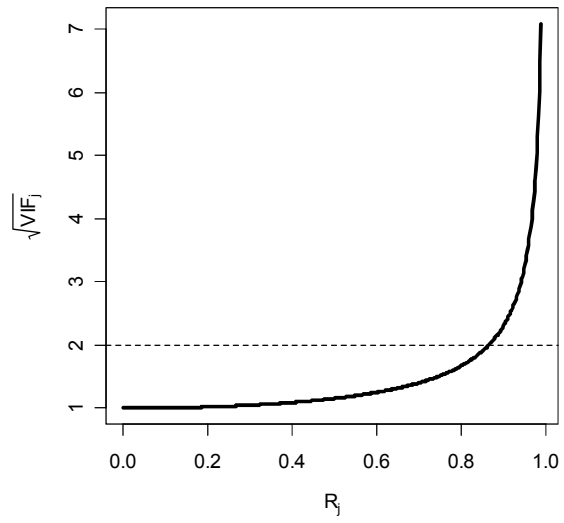

Figure 2. The square root of the variance-inflation factor as a function of the multiple correlation for the regression of X_j on the other X's.

3.1 Generalized Variance-Inflation Factors

- Variance-inflation factors are not fully applicable to models that include related sets of regressors, such as dummy regressors constructed from a polytomous categorical variable or polynomial regressors.
- The reasoning underlying this qualification is subtle:
  - The correlations among a set of dummy regressors are affected by the choice of baseline category and coding scheme.
  - Similarly, the correlations among a set of polynomial regressors in an explanatory variable X (e.g., X, X^2, X^3) are affected by adding a constant to the X-values.
  - Neither of these changes alters the fit of the model to the data, however, so neither is fundamental.
  - We are not concerned, therefore, with the 'artificial' collinearity among dummy regressors or polynomial regressors in the same set.

  - We are instead interested in the relationships between regressors generated to represent the effects of *different* explanatory variables.
  - As a consequence, we can legitimately employ variance-inflation factors to examine the impact of collinearity on the coefficients of numerical regressors, or on any single-degree-of-freedom effects (such as a single dummy variable constructed to represent a dichotomous explanatory variable), even when sets of dummy regressors or polynomial regressors are present in the model.
- Fox and Monette generalize the notion of variance inflation to sets of related regressors.
  - The standard variance-inflation factors represents the impact of collinearity on the squared length of the confidence interval for a coefficient.
  - The generalization of a confidence interval for a single coefficient is a *joint confidence region* for two or more coefficients.

  - Fox and Monette show how to measure the impact of collinearity on the squared size of the confidence region for p coefficients, calling the resulting index a *generalized variance-inflation factor*.
    - They suggest taking the $1/2p$ power of the GVIF to make the index comparable across different values of p.
    - This is analogous to using the square root of the VIF for a single coefficient.
- Suppose that the regressors of interest are collected into a set called $\mathbf{X}_1$, and the remaining regressors are in $\mathbf{X}_2$.
  - The set $\mathbf{X}_1$, for example, could contain related dummy regressors.
  - Then

$$\text{GVIF}_1 = \frac{\det \mathbf{R}_{11} \det \mathbf{R}_{22}}{\det \mathbf{R}}$$

  - Here, $\mathbf{R}_{11}$ is the correlation matrix for $\mathbf{X}_1$;
  - $\mathbf{R}_{22}$ is the correlation matrix for $\mathbf{X}_2$;
  - and $\mathbf{R}$ is the matrix of correlations among all of the variables.
  - $\det$ is a number associated with a square matrix, called its *determinant*. Determinants are a standard topic in linear algebra.

- The GVIF does not depend on non-essentials such as choice of baseline category or coding scheme for dummy regressors.

4. Coping With Collinearity: No Quick Fix

- When X_1 and X_2 are strongly collinear, the data contain little information about the impact of X_1 on Y holding X_2 constant, because there is little variation in X_1 when X_2 is fixed. (Think about the added-variable plot for X_1.)
  - Of course, the same is true for X_2 fixing X_1.
- There are several strategies for dealing with collinear data but none magically extracts nonexistent information from the data.
  - Rather, the research problem is redefined, often subtly and implicitly.
  - Sometimes the redefinition is reasonable; usually it is not.
- The ideal solution to the problem of collinearity is to collect new data in such a manner that the problem is avoided — for example, by experimental manipulation of the X's.
  - This solution is rarely practical.
- There are several less adequate solutions:

4.1 Model Re-Specification

- Although collinearity is a data problem, not (necessarily) a deficiency of the model, one approach is to re-specify the model.
  - Perhaps several regressors in the model can be conceptualized as alternative indicators of the same construct.
    - Then these measures can be combined or one can be chosen to represent the others.
    - High correlations among the X's indicate high reliability.
  - Alternatively, we can reconsider whether we really need to control for X_2 (for example) in examining the relationship of Y to X_1.
    - Re-specification of this variety is possible only where the original model was poorly thought out, or where the researcher is willing to abandon (some of) the goals of the research.

4.2 Variable Selection

- A common, but usually misguided, approach to collinearity is variable selection, where some procedure is employed to reduce the regressors in the model to a less highly correlated set.
- *Stepwise Methods:*

(1) Forward stepwise methods add explanatory variables to the model one at a time. At each step, the variable that yields the largest increment in R^2 is selected. The procedure stops when the increment is smaller than a preset criterion.

(2) Backward stepwise methods are similar, except that the procedure starts with the full model and deletes variables one at a time.

(3) Forward/backward methods combine the two approaches.

- Stepwise methods frequently are abused by researchers who interpret the order of entry of X's as an index of their 'importance.'
  - Suppose that there are two highly correlated X's that have nearly identical large correlations with Y; only one X will enter the regression equation.
  - A small modification to the data, or a new sample, could easily reverse the result.

▶ *Subset Methods:*

- Stepwise methods can fail to turn up the optimal subset of regressors of a given size.
- It is feasible to examine *all* subsets of regressors even when k is large.
- Subset techniques also have the advantage of revealing alternative, nearly equivalent models, and thus avoid the appearance of a uniquely 'correct' result.

▶ Some additional cautions about variable selection:

- Variable selection results in a re-specified model that usually does not address the original research questions.
  - If the original model is correctly specified, then coefficient estimates following variable selection are biased.
- When regressors occur in sets (e.g., of dummy variables), then these sets should be kept together during selection.
  - Likewise, when there are hierarchical relations among regressors, these relations should be respected — for example, don't remove a main effect when an interaction to which it is marginal is included in the model.
- Coefficient standard errors calculated following explanatory-variable selection overstate the precision of results.

▶ Variable selection has applications to statistical modeling even when collinearity is not an issue, particularly in prediction (see below).

4.3 Biased Estimation

▶ The essential idea here is to trade a small amount of bias in the coefficient estimates for a substantial reduction in coefficient sampling variance, producing a smaller mean-squared error of estimation of the β's.

▶ By far the most common biased estimation method is *ridge regression* (due to Hoerl and Kennard).

▶ Like variable selection, biased estimation is not a panacea for collinearity.

- Ridge regression involves the arbitrary selection of a 'ridge constant,' which controls the extent to which ridge estimates differ from the least-squares estimates.
- The larger the ridge constant, the greater the bias and the smaller the variance of the ridge estimator (see Figure 3).
- To pick an optimal ridge constant — or even a good one — generally requires knowledge about the unknown β's.

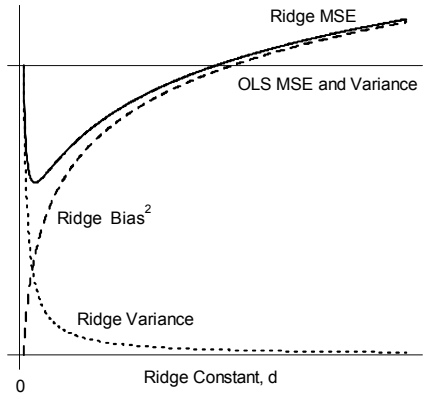

Figure 3. The squared bias, variance, and MSE of the ridge estimator as a function of the ridge constant, compared to the MSE of the OLS estimator.

4.4 Prior Information About the Regression Coefficients

- A final approach is to introduce additional prior information that reduces the ambiguity produced by collinearity.
- Here is a particularly simple case:
  - We wish to estimate the model
    $$Y = \alpha + \beta_1 X_1 + \beta_2 X_2 + \beta_3 X_3 + \varepsilon$$
    where Y is savings, X_1 is income from wages and salaries, X_2 is dividend income from stocks, and X_3 is interest income.
  - We have trouble estimating β_2 and β_3 because X_2 and X_3 are highly correlated.
  - We have reason to believe that $\beta_2 = \beta_3$, and denote the common quantity β_*.
  - If X_2 and X_3 were not so highly correlated, then we could test this belief as a hypothesis.

- In the current situation, we can fit the model
  $$Y = \alpha + \beta_1 X_1 + \beta_*(X_2 + X_3) + \varepsilon$$
  incorporating our belief in the equality of β_2 and β_3 in the specification of the model, and thus eliminating the collinearity problem (along with the possibility of testing the belief).

4.5 Some Comparisons

- The several approaches to collinear data have much in common:
  - Model re-specification can involve variable selection, and variable selection re-specifies the model.
  - Variable selection constrains the coefficients of deleted regressors to zero.
  - Variable selection produces biased coefficient estimates.
    - We hope that the trade off of bias against variance is favorable, but because the bias depends on the unknown regression coefficients, we have no assurance that this will be the case.
  - Certain types of prior information result in a re-specified model.
  - Biased-estimation methods like ridge regression place prior constraints on the values of the β's.

- *Conclusion*: Mechanical model-selection and modification procedures disguise the substantive implications of modeling decisions.
  - These methods generally cannot compensate for weaknesses in the data and are no substitute for judgment and thought.

5. Model Selection

- I have touched on issues of model selection at several points, often simplifying a model after preliminary statistical hypothesis tests
- Issues of model search extend beyond the selection of explanatory variables or terms to include in a regression model to questions such as the removal of outliers and variable transformations.
- The strategy of basing model selection on hypothesis tests is problematic for a number of reasons:
  - Simultaneous inference.
  - The fallacy of affirming the consequent.
  - The impact of large samples on hypothesis tests.
  - Exaggerated precision following model selection.

- There are several general strategies for addressing these concerns:
  - Using alternative model-selection criteria.
  - Compensating for simultaneous inference (e.g., by Bonferroni adjustment or by model validation).
  - Avoiding model selection by resisting the temptation of simplify a model.
  - Model averaging — accounting for uncertainty by weighting alternative models according to their degree of support from the data.
  - Mode validation — using part of the data to develop a statistical model and the rest for statistical inference.

5.1 Model Selection Criteria

- Model selection is conceptually simplest when our goal is *prediction*—that is, the development of a regression model that will predict new data as accurately as possible.
- I assume that we have n observations on a response variable Y, and a set of m contending statistical models $\mathcal{M} = \{M_1, M_2, \ldots, M_m\}$ for Y; model M_j has s_j regression coefficients.
- R^2 "corrected" for degrees of freedom:

$$\begin{aligned}\widetilde{R}_j^2 &\equiv 1 - \frac{S_E^{(j)2}}{S_Y^2} \\ &= 1 - \frac{n-1}{n-s_j} \times \frac{\mathsf{RSS}_j}{\mathsf{TSS}}\end{aligned}$$

  where RSS_j is the residual sum of squares and $S_E^{(j)2}$ the residual variance under model M_j.

- Mallows's C_p statistic, which estimates the mean-squared error of prediction under the model:

$$\begin{aligned}C_{p_j} &\equiv \frac{\sum E_i^{(j)2}}{S_E^2} + 2s_j - n \\ &= (k+1-s_j)(F_j - 1) + s_j\end{aligned}$$

  where the error variance estimate S_E^2 is based on the *full* model fit to the data, containing all $k+1$ regressors; and F_j is the incremental F-statistic for testing the hypothesis that the regressors omitted from model M_j have population coefficients of 0.
- The cross-validation criterion:

$$\mathsf{CV}_j \equiv \frac{\sum_{i=1}^n \left(\widehat{Y}_{-i}^{(j)} - Y_i\right)^2}{n}$$

  where $\widehat{Y}_{-i}^{(j)}$ is the predicted value for the ith obseration obtained from a model fit without this observation. We prefer the model with the smallest value of CV_j.

- The generalized cross-validation criterion:
$$\text{GCV}_j \equiv \frac{n \times \text{RSS}_j}{df^2_{\text{res}_j}}$$
- The Akaike information criterion:
$$\text{AIC}_j \equiv n \log_e \widehat{\sigma}^{(j)2}_\varepsilon + 2s_j$$
where $\widehat{\sigma}^{(j)2}_\varepsilon = \left(\sum E^{(j)2}_i\right)/n$ is the MLE of the error variance for model j.
- Schwartz's Bayesian information criterion:
$$\text{BIC}_j \equiv n \log_e \widehat{\sigma}^{(j)2}_\varepsilon + s_j \log_e n$$
Note that the BIC penalizes lack of parsimony more than the AIC does.
- Regardless of the criterion applied, automatic model-selection methods attend to the predictive adequacy of regression models and are blind to their substantive interpretability.

- Model-selection criteria such as the AIC and BIC are not limited to comparing models selected by automatic methods.
  - One of the currently popular applications of the BIC is to justify the removal of small, but "statistically significant," terms in regression models fit to large samples of data.
  - But researchers should feel free to remove "statistically significant" terms from a statistical model based on the substantive judgment that these terms are too small to be of interest *regardless* of the value of the BIC.

5.2 Model Validation

- In *model validation*, part of the data (called the "training" or "exploratory" sub-sample) is used to specify a statistical model, which is then evaluated using the other part of the data (the "validation" or "confirmatory" sub-sample).
- The process of data exploration, model fitting, model criticism, and model re-specification is typically iterative, requiring several failed attempts before an adequate description of the data is achieved.
  - In the process, variables may be dropped from the model; terms such as interactions may be incorporated or deleted; variables may be transformed; and unusual data may be corrected, removed, or otherwise accommodated.
  - The risk of iterative modeling is that we will capitalize on chance—over-fitting the data and overstating the strength of our results.

- An ideal solution would be to collect new data with which to validate a model, but this solution is often impractical.
- Model validation simulates the collection of new data by randomly dividing the data that we have in hand into two parts—the first part to be used for exploration and model formulation, the second for checking the model, formal estimation, and testing.
- Barnard: "The simple idea of splitting a sample into two and then developing the hypothesis on the basis of one part and testing it on the remainder may perhaps be said to be one of the most seriously neglected ideas in statistics, if we measure the degree of neglect by the ratio of the number of cases where a method could give help to the number where it is actually used."

6. Summary

- Perfect collinearity occurs when one regressor in a linear model is a perfect linear function of others.
  - Under prefect collinearity, the least-squares regression coefficients are not unique.
- Less-than-perfect collinearity occurs when one regressor is highly correlated with others, a situation that causes its regression coefficient to become unstable
  - For example, the standard error of the coefficient is much larger than it would otherwise be.
- The variance-inflation factor $1/(1-R_j^2)$ indicates the impact of collinearity on the precision of B_j.
  - Variance inflation can be extended to sets of regression coefficients, such as coefficients for a set of related dummy regressors.

- Several methods are employed to deal with collinearity problems (short of collecting new, non-collinear data), including model respecification, variable selection, biased estimation, and the introduction of additional information.
  - The first three approaches have intrinsic limitations, and the fourth is rarely practical.
- There are several criteria beyond hypothesis testing that can be used for model selection.
  - Automatic methods of model selection are justified for pure prediction problems and are otherwise problematic.
- Model validation protects the integrity of statistical inference following model selection by dividing the data into two parts — a training subsample used to develop a statistical model and a validation subsample used to check the model and to perform statistical inference.