Collinearity and Model Selection

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.

Collinearity

When there is a perfect linear relationship among the regressors in a linear model, the least-squares coefficients are not uniquely defined.

When there is a perfect linear relationship among the X's

– 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.

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 wide.

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.

– Insufficient variation in explanatory variables, small samples, and large error variance can also result in imprecise estimates.

Variance Inflation Factors (VIFs)

The sampling variance of the least-squares slope coefficient $\widehat{\beta}_j$ is:

$$V(\widehat{\beta}_j) = \frac{1}{1 - R_j^2} \times \frac{\sigma_\epsilon^2}{(n-1)S_j^2}$$

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 $\widehat{\beta}_j$.

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 0.9 that the precision of estimation is halved.

VIFs are not fully applicable to models that include related sets of regressors, such as dummy regressors constructed from a variable with multiple categories 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 employ VIFs 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 binary predictor variable), even when sets of dummy regressors or polynomial regressors are present in the model.

The notion of VIFs can be generalized to sets of related regressors, and these are called generalized VIFs (GVIFs).

Solving a Multicollinearity Problem

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.

Of course, the same is true for X_2, holding constant X_1.

There are several strategies for dealing with collinear data but none magically extracts nonexistent information from the data.

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 — but this solution is rarely practical.

There are several less adequate solutions that often subtly and implicitly redefine the research problem, which may or may not be a reasonable redefinition.

Model Respecification 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.

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:

(a) 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 (or some other criteria) is selected. The procedure stops when the increment is smaller than a preset criterion.

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

(c) 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 question.
- 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, do not remove a main effect when an interaction to which it is marginal is included in the model.
- Coefficient standard errors calculated following variable selection overstate the precision of results.

Variable selection has applications to statistical modeling even when collinearity is not an issue, particularly in prediction.

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.

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.

- To pick an optimal ridge constant — or even a good one — generally requires knowledge about the unknown β's.

Summary

Perfect collinearity occurs when one regressor in a linear model is a perfect linear function of others.

Under perfect 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 (VIF), $1/(1 - R_j^2)$, indicates the impact of collinearity on the precision of $\widehat{\beta}_j$.

VIFs 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, and biased estimation.

Model Selection

We have compared the fit of two nested models using the `anova()` function. This method does not work if the two models are not not nested. A *nested model* is one whose terms are completely included in the other model.

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.

One way to automatically select a model is to begin with the largest model you can, and then prune it, which can be done in several ways:

- Eliminate the least-significant coefficient.
- Pick your favorite model selection criterion, consider deleting each coefficient in turn, and pick the sub-model with the best value of the criterion.

Having eliminated a variable, one then re-estimates the model, and repeats the procedure. Stop when either all the remaining coefficients are significant (under the first option), or nothing can be eliminated without worsening the criterion.

This is **backwards** stepwise model selection. **Forward** stepwise model selection starts with the intercept-only model and adds variables in the same fashion. There are also forward-backward hybrids.

Stepwise model selection is a **greedy** procedure: it takes the move which does the most to immediately improve the criterion, without considering the consequences down the line. There are very, very few situations where it is consistent for model selection, or (in its significance testing version) where it even does a particularly good job of coming up with predictive models, but it is surprisingly popular.

Regardless of the criterion applied, automatic model selection methods attend to the predictive adequacy of regression models and are blind to their substantive interpretability.

The strategy of basing model selection on signficance testing 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., Bonferroni adjustment).
- Avoiding model selection by resisting the temptation to simplify a model.
- Model averaging — accounting for uncertainty by weighting alternative models according to their degree of support from the data.
- Model validation — using part of the data to develop a statistical model and the remaining data for statistical inference.

Adjusted $\widetilde{R}^2$

Recall that R^2 is:

$$R^2 \equiv 1 - \frac{RSS}{TSS}$$

But this value can only increase as predictors or interaction terms are added. So adjusted $\widetilde{R}^2$ penalizes the R^2 value by a "correction" for the degrees of freedom.

Adjusted $\widetilde{R}^2$:

$$\widetilde{R}^2_j \equiv 1 - \frac{S^2_\epsilon}{S^2_Y} = 1 - \frac{\frac{RSS}{n-k-1}}{\frac{TSS}{n-1}} = 1 - \frac{n-1}{n-k-1} \times \frac{RSS}{TSS}$$

where S^2_ϵ the residual variance.

Unless the sample size is small, R^2 and $\widetilde{R}^2$ are usually similar.

Mallows's C_p

Assume that we have n observations on a response variable Y, and a set of m contending statistical models $M_1, M_2, \ldots, M_m$ for Y; model M_j has s_j regression coefficients.

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

$$C_{pj} \equiv \frac{\sum \epsilon^{(j)2}_i}{S^2_\epsilon} + 2s_j - n = (k + 1 + s_j)(F_j - 1) + s_j$$

where the error variance estimate S^2_ϵ 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.

Akaike information criterion (AIC)

The Akaike Information Criterion (AIC) provides another method for comparing models. The index takes into account a model's statistical fit and the number of parameters needed to achieve this fit. Models with smaller AIC values—indicating adequate fit with fewer parameters—are preferred.

The Akaike information criterion:

$$AIC_j \equiv n \log_e \widehat{\sigma}^{(j)2}_\epsilon + 2s_j$$

where $\widehat{\sigma}^{(j)2}_\epsilon = \left(\sum \epsilon^{(j)2}_i\right)/n$ is the estimate of the error variance for model M_j.

Bayesian information criterion (BIC)

Schwartz's Bayesian information criterion

$$BIC_j \equiv n \log_e \widehat{\sigma}_\epsilon^{(j)2} + s_j \log_e n$$

Note that the BIC penalizes lack of parsimony more than the AIC does.

As $n \to \infty$, if the true model is among those BIC can select among, BIC will tend to pick the true model. Unfortunately, the model selected by BIC will tend to predict less well than the one selected by AIC or leave-one-out cross-validation.

Model selection criteria such as the AIC and BIC are not limited to comparing models selected by automatic methods (e.g., stepwise).

Leave-one-out-cross-validation

When looking at influential observations and outliers, we considered omitting one observation from the data set, estimating the model, and then trying to predict the outcome for that observation. The **leave-one-out** fitted value for observation i is $\widehat{Y}_{-i}^{(j)}$, where the subscript $(-i)$ indicates that observation i was left out in calculating this fit. The **leave-one-out cross-validation criterion** for model j is

$$CV_j = \frac{\sum_{i=1}^{n} (\widehat{Y}_{-i}^{(j)} - Y_i)^2}{n}$$

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

The idea is that we want to know if our model can generalize to new data, so we *see* how well it generalizes to new data. Leaving out each observation in turn ensures that that the set of observations on which we try to make predictions is just as representative of the whole population as the original sample was. Leave-one-out CV is an unbiased estimate of this generalization error.

Short-cut Based on Leverage Re-estimating the model n times would be seriously time-consuming, but there is fortunately a short-cut:

$$CV = \frac{1}{n} \sum_{i=1}^{n} \left(\frac{Y_i - \widehat{m}_i}{1 - H_{ii}} \right)^2$$

The numerator inside the square is the residual of the model fit to the full data. This gets divided by $1 - H_{ii}$, which is also something we can calculate with just one fit to the model. The denominator says that the residuals for high-leverage observations count more, and those for low-leverage observations count less.

k-Fold Cross-Validation

In leave-one-out CV, we omitted each observation in turn, and tried to predict it. k-fold CV is somewhat different, and goes as follows.

- Randomly divide the data into k equally-sized parts, or "folds".
- For each fold
  - Temporarily hold back that fold, calling it the "testing set".
  - Call the other $k - 1$ folds, taken together, the "training set".
  - Estimate each model on the training set.
  - Calculate the MSE of each model on the testing set.
  - Average MSEs over the k folds.

We then pick the model with the lowest MSE, averaged across testing sets.

The idea is just like leave-one-out CV: the models are compared only on data which they did not get to "see"" during estimation. Indeed, leave-one-out CV is a special case of k-fold CV where $k = n$. The disadvantage of doing leave-one-out CV, is that all of the training sets are very similar (they share $n - 2$ data points), so averaging over folds does very little to reduce variance. For moderate k — typically 5 or 10 — k-fold CV tends to produce very good model selection results.

The problem with multiple hypothesis signficance testing

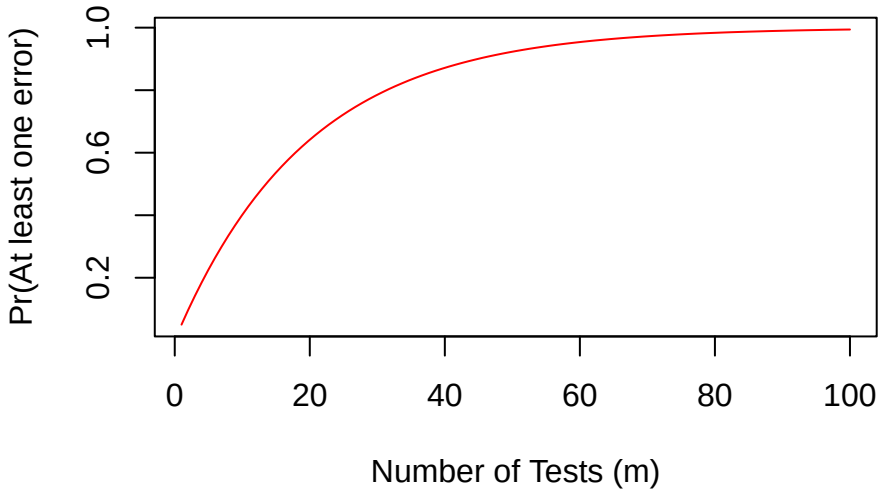
When we test any statistical hypothesis, our decision may be wrong.

We use a procedure to control the probability of false rejection at α, often $\alpha = .05$.

If our data analysis involves many hypothesis tests, the probability of at least one Type I error increases rather sharply with the number of tests.

For example, if there are m tests and they are independent, and each one is performed with a Type I error rate of α, and all hypotheses are actually true, the probability of at least one Type I error is $1 - P(\text{No errors}) = 1 - (1 - \alpha)^m$.

```
curve(1 - (1 - 0.05)^x, 1, 100, xlab = "Number of Tests (m)", ylab = "Pr(At least one error)", col = "r
```



A key decision in analyzing data is to decide on the set of hypotheses to consider as a family.

A family is a set for which signi

cance statements and related error rates will be controlled jointly.

The term "experiment" is sometimes used instead of "family."

Even within the same data set, different families may be analyzed for different reasons. For example, suppose you have data for 50 hospitals. You may be interested in all the pairwise comparisons among the hospitals. On the other hand, the hospital administrators at hospital A may only be interested in the family of pairwise comparison of their hospital with the other 49.

The *error rate per hypothesis*, often called the error rate per comparison or PCER, is the Type I error rate for each individual hypothesis test.

The *error rate per family*, or PFER, is the expected number of false rejections in the family.

The *familywise error rate* (FWER) is the probability of at least one Type I error in the family of tests.

Let V_m be the number of Type I errors committed in a family of tests, and R_m be the number of rejected hypotheses.

The generalized familywise error rate is, $gFWER(k) = Pr(V_m > k)$, the chance of at least $(k+1)$ false positives. The special case $k = 0$ corresponds to the usual FWER.

The *False Discovery Rate* (FDR) is (V_m/R_m), the long run proportion of rejections that are Type I errors.

Bonferroni method to control FWER A finite set of minimal hypotheses $H_i, i = 1, \ldots, m$ is to be tested. Corresponding to the H_i are test statistics T_i (or their absolute values) such that p_i-values corresponding to each hypothesis may be computed.

Assume that the p_i-values are ordered such that $p_1 \leq p_2 \leq \ldots \leq p_m$.

A simple method is to reject H_i if $p_i \leq \alpha_i$ where $\sum_{i=1}^{m} \alpha_i = \alpha$.

This method controls FWER at or below α.

Usually, all α_i are set equal to α/m.

This correction is too conservative. Power suffers increasingly as m becomes large. FDR is a less stringent condition than the FWER.

False Discovery Rate Classical procedures that control the FWER at levels conventional in single-comparison problems, tend to have substantially less power than the per comparison procedure of the same levels.

FDR is the expected proportion of the rejected tests that should not have been rejected (i.e., the expected proportion of false discoveries amongst the rejected hypotheses).

The Benjamini-Hochberg method is as follows: Suppose there are m null hypotheses, and, unknown to the data analyst, m_0 are true. The following method controls FDR at or below $\alpha m_0/m$ (which is $\leq \alpha$).

Consider again the ordered p_i-values, $p_1 \leq p_2 \leq \ldots \leq p_m$, reject the set of hypotheses $H_1, H_2, \ldots, H_k$ for which

$$k = \max\left(i : p_i \leq \frac{i}{m}\alpha\right)$$

The `p.adjust()` function in R takes as its argument a vector of p-values. The `method=` choices are `"holm"`, `"hochberg"`, `"hommel"`, `"bonferroni"`, `"BH"`, `"BY"`, and `"fdr"`, which is the same as `"BH"`. Benjamini-Hochberg (`"BH"`) and Benjamini-Yekutieli (`"BY"`) are both FDR corrections. The `"holm"`, `"hochberg"`, `"hommel"`, and `"bonferroni"` methods all control the FWER. `"BH"` was designed for when p-values are independent. If they are not, then use `"BY"`.

```
pvals <- c(.001, .001, .001, .02, .22, .59, .87)
BH <- p.adjust(pvals , method="BH")
Bon <- p.adjust(pvals , method="bonferroni")
results <- cbind(pvals, BH=round(BH, 3), Bon=round(Bon, 3))
results
```

```
##      pvals    BH   Bon
## [1,] 0.001 0.002 0.007
## [2,] 0.001 0.002 0.007
## [3,] 0.001 0.002 0.007
## [4,] 0.020 0.035 0.140
## [5,] 0.220 0.308 1.000
## [6,] 0.590 0.688 1.000
## [7,] 0.870 0.870 1.000
```

Inference after Selection

All of the inferential statistics we have used presumed that our choice of model was completely fixed, and not at all dependent on the data. If different data sets would lead us to use different models, and our data are (partly) random, then which model we are using is also random. This leads to some extra uncertainty in, for example, our estimate of the slope on X_1, that is *not* accounted for by our formulas for the sampling distributions, hypothesis tests, confidence intervals, etc.

For example, I simulate 200 data points where the Y variable is a standard normal variate, and there are 100 independent predictor variables, all also standard normal variables, independent of each other *and of* Y:

```
set.seed(1234)
n <- 200
k <- 100
y <- rnorm(n)
x <- matrix(rnorm(n*k), nrow=n)
df <- data.frame(y=y, x)
mdl <- lm(y~., data=df)
```

Of the 100 predictors, 7 have t-statistics which are significant at the 0.05 level or less. (The expected number would be 5.) If I select the model using just those variables, I get the following:

```
stars <- 1 + which(coefficients(summary(mdl))[-1,4] < 0.05)
mdl.2 <- lm(y~., data=df[,c(1,stars)])
summary(mdl.2)
```

```
##
## Call:
## lm(formula = y ~ ., data = df[, c(1, stars)])
##
## Residuals:
##      Min       1Q   Median       3Q      Max
## -2.93248 -0.59574 -0.09111  0.54567  2.90490
##
## Coefficients:
##             Estimate Std. Error t value Pr(>|t|)
## (Intercept) -0.04800    0.06851  -0.701 0.484387
## X4          -0.21205    0.06955  -3.049 0.002621 **
## X12          0.15314    0.07179   2.133 0.034173 *
## X13         -0.25555    0.06640  -3.848 0.000162 ***
## X26         -0.06919    0.06849  -1.010 0.313662
## X67          0.08660    0.07202   1.202 0.230710
## X78         -0.15689    0.06922  -2.267 0.024528 *
## X85         -0.10430    0.06299  -1.656 0.099375 .
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 0.9527 on 192 degrees of freedom
## Multiple R-squared:  0.1594, Adjusted R-squared:  0.1288
## F-statistic: 5.203 on 7 and 192 DF,  p-value: 1.898e-05
```

Notice that the final overall F statistic testing whether including those variables fits better than an intercept-only model has a significant p-value. This is the case even though, by construction, the response is *completely independent* of *all* predictors. This is not a fluke: if we re-run this simulation many times, the confidence intervals would also be much too narrow.

These issues do not go away if the true model is not "everything is independent of everything else", but rather

has some structure. Because we picked the model to predict well on this data, if we then run hypothesis tests on that same data, they will be too likely to tell us everything is significant, and our confidence intervals will be too narrow. Thus, doing statistical inference on the same data we used to select the model is generally inappropriate.

A solution is to use *different data sets* to select a model and to do inference with the selected model.

Data Splitting Data splitting is a very simple procedure:

1. Randomly divide your data set into two parts.

2. Calculate your favorite model selection criterion for all your candidate models using only the first part of the data. Pick one model as the winner.

3. Re-estimate the winner, and calculate all your inferential statistics, using only the other half of the data.

Division into two equal halves is optional, but usual.

Because the winning model is statistically independent of the second half of the data, we can treat it as though that model were fixed *a priori*. Since we are only using $n/2$ data points to calculate confidence intervals, etc., they will be somewhat wider than if we really had fixed the model in advance and used all n data points, but that is the price we pay for having to select a model based on data.

Summary

There are several criteria beyond signficance 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 splitting the data into two parts — a subsample used to develop a statistical model and a validation subsample used to check the model and to perform statistical inference.