Regression Diagnostics

Diagnosing Collinearity

last modified: 2022-02-08

John Fox

McMaster University

SORA/TABA 2022

Outline

1. Collinearity Diagnostics
2. Measuring Collinearity: Variance Inflation
3. Visualizing Collinearity
4. Generalized Variance Inflation
5. Dealing with Collinearity: No Quick Fix
6. References

Outline

1. Collinearity Diagnostics
2. Measuring Collinearity: Variance Inflation
3. Visualizing Collinearity
4. Generalized Variance Inflation
5. Dealing with Collinearity: No Quick Fix
6. References

Collinearity Diagnostics

- Collinearity is different from the other problems I have discussed:
  - Except in exceptional circumstances, collinearity is fundamentally a problem with the data rather than with the specification of the regression model.
  - Consequently, there is usually no satisfactory solution for a true collinearity problem.
- I'll first address collinearity diagnostics for linear least-squares regression and then generalize to other regression models.
- In least-squares regression, *perfect* collinearity implies that the model matrix $\mathsf{X}_{n \times k+1}$ is of rank $r < k+1$.
  - This implies that $\mathsf{X}^\mathsf{T}\mathsf{X}$ is also of rank $r < k+1$ and therefore is singular; as a consequence, the least-squares regression coefficients, usually $\mathsf{b} = (\mathsf{X}^\mathsf{T}\mathsf{X})^{-1}\mathsf{X}^\mathsf{T}\mathsf{y}$, are not unique.
  - Perfect collinearity usually reflects a bone-headed error, such as including p rather than $p-1$ dummy regressors, along with an intercept, to represent a factor with p levels, or fitting a model to data for which there are more coefficients than cases, $k+1 > n$.

Outline

1. Collinearity Diagnostics
2. Measuring Collinearity: Variance Inflation
3. Visualizing Collinearity
4. Generalized Variance Inflation
5. Dealing with Collinearity: No Quick Fix
6. References

Measuring Collinearity: Variance Inflation

- As I noted much earlier, the estimated sampling variance of the least-squares regression coefficient b_j is $\widehat{V}(b_j) = \dfrac{s^2}{(n-1)s_j^2} \times \dfrac{1}{1-R_j^2}$, where where R_j^2 is the squared multiple correlation from the regression of x_j on the other xs.
  - The impact of collinearity on the precision of estimation is captured by $1/(1-R_j^2)$, called the *variance-inflation factor* VIF_j, implicitly comparing the data at hand to similar "utopian" data with uncorrelated xs.
  - The other factors affecting the variance of b_j are the estimated error variance s^2, the sample size n, and the variance s_j^2 of x_j.
  - When estimated coefficients are imprecise, the culprits are much more likely to be weak explanatory variables, too-small samples, and homogeneous xs than collinearity.
  - The square-root of VIF_j expresses the effect of collinearity on the standard error of b_j and hence on the width of a confidence interval for β_j; it's not until $R_j \approx .9$ that $\sqrt{\mathrm{VIF}_j} > 2$.

Outline

1. Collinearity Diagnostics

2. Measuring Collinearity: Variance Inflation

3. Visualizing Collinearity
   - Data and Confidence Ellipses
   - Added-Variable Plots

4. Generalized Variance Inflation

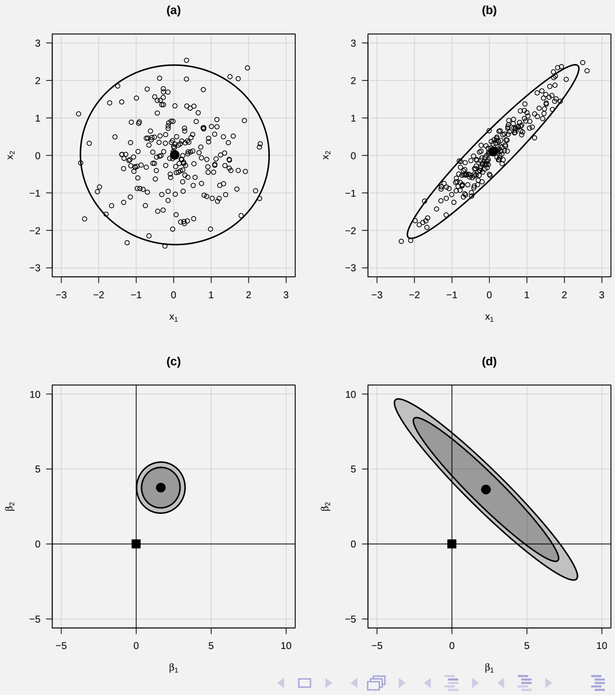
5. Dealing with Collinearity: No Quick Fix

6. References

Visualizing Collinearity

Data and Confidence Ellipses

- 95% *data ellipses* for the regressors x_1 and x_2 and confidence ellipses for β_1 and β_2 in the regression of y on x_1 and x_2.
  1. $n = 200$ values of x_1 and x_2 were sampled from a bivariate normal distribution with $\mu_1 = \mu_2 = 0$ and $\sigma_1 = \sigma_2 = 1$. For (a), the correlation $\rho_{12} = 0$, while for (b), $\rho_{12} = 0.95$. The sample correlations are $r_{12} = -0.007$ and $r_{12} = 0.957$.
  2. The response was generated according to $y = 2x_1 + 3x_2 + \varepsilon$ and $\varepsilon \sim \mathrm{NID}(0, 10^2)$.
  - The outer confidence ellipses are at the 95% level and the projections of the inner ellipses onto the β_1 and β_2 axes produce 95% confidence *intervals*.



Visualizing Collinearity

Data and Confidence Ellipses

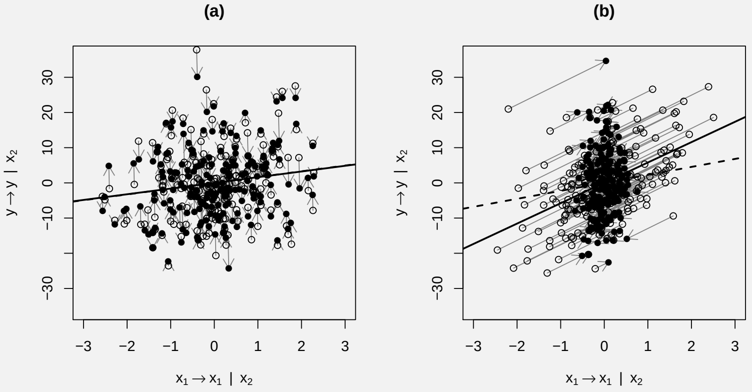
- The confidence ellipse is a 90° rotation and rescaling of the data ellipse.
- Here are summaries for the regressions in the two artificial data sets:

	(a) $r_{12} = -0.007$		*(b)* $r_{12} = 0.957$	
	Estimate	*SE(b)*	*Estimate*	*SE(b)*
b_0	−0.087	0.688	−0.177	0.672
b_1	1.642	0.656	2.270	2.480
b_2	3.760	0.690	3.630	2.450
s	9.44		9.44	
R^2	0.154		0.252	
$\sqrt{\text{VIF}}$	1.00		3.45	

Visualizing Collinearity

Added-Variable Plots

- The graphs at the right superimpose the AV plot for x_1 on the marginal scatterplot for y and x_1 in the regressions of y on x_1 and x_2 in the two artificial data sets, (a) where $r_{12} = -0.007$, and (b) where $r_{12} = 0.956$
- The arrows show the correspondence between points in the marginal (open circles) and added-variable (filled circles) plots.
- The solid line is the least-squares line for the marginal scatterplot and the broken line is the least-squares line for the added-variable plot, giving the multiple-regression slope b_1.



- In (b) the conditional variation of $x_1|x_2$ in the AV plot is considerably reduced from the marginal variation of x_1, while in (a) only the residual variation is reduced in the AV plot.

Outline

1. Collinearity Diagnostics
2. Measuring Collinearity: Variance Inflation
3. Visualizing Collinearity
4. Generalized Variance Inflation
5. Dealing with Collinearity: No Quick Fix
6. References

Generalized Variance Inflation

- The VIF is only a sensible measure for terms in a model that are represented by a single parameter.
  - Multiple-parameter terms include sets of dummy-regressor coefficients for factors with more than two levels, and polynomial or regression-spline coefficients for numeric explanatory variables.
  - For example, correlations among a set of dummy regressors depend on which level is selected as the baseline level, but the fit of the model to the data and intrinsic meaning of the model don't change with this arbitrary choice.

Generalized Variance Inflation

- Fox and Monette (1992) introduce *generalized variance-inflation factors* (*GVIFs*) to deal with sets of related regression coefficients.
    - The GVIF for two coefficients (say, for two dummy regressors) is the increase in the squared *area* of the joint-confidence ellipse for the two corresponding parameters, relative to the area of this ellipse for otherwise similar data in which the two regressors are unrelated to the *other* regressors in the model.
    - This ratio of squared areas is unaffected by the choice of baseline level for the set of dummy regressors or by other similar arbitrary choices.
    - If there are three coefficients in a set, then the GVIF represents inflation in the squared *volume* of the joint-confidence ellipsoid for the coefficients, and the generalization beyond three coefficients is to the squared *hypervolume* of the multidimensional confidence ellipsoid for the coefficients.
    - Because the size of the GVIF tends to grow with the number of regressors p in a set, Fox and Monette recommend taking the $2p$th root of the GVIF, i.e., $\mathrm{GVIF}^{\frac{1}{2p}}$.
    - When $p = 1$, the GVIF reduces to the usual VIF.

Generalized Variance Inflation

- To compute the GVIF, partition the regressors in the regression models into two sets: (1) those for the term in question (e.g., the set of dummy regressors for a factor), and (2) the remaining regressors in the model, with the exception of the constant regressor $x_0 = 1$, which is ignored.
- Let R represent the correlation matrix among all of the regressors (again, ignoring the constant), R_1 the correlations among the regressors in the first set, and R_2 the correlations among the regressors in the second set.
- Then the generalized variance-inflation factor for the first (or indeed second) set of regressors is $\mathrm{GVIF}_1 = \det(\mathrm{R}_1)\det(\mathrm{R}_2)/\det(\mathrm{R})$.

Generalized Variance Inflation

- We can also express the GVIF in terms of the correlation matrix R_b of the regression coefficients, computed from $\widehat{V}(\mathrm{b}) = s^2(\mathrm{X}^T\mathrm{X})^{-1}$ after eliminating the first row and column for the intercept.
  - Let R_{b_1} be the submatrix of R_b pertaining to the correlations of the coefficients in set 1 and R_{b_2} the submatrix pertaining to the correlations of the coefficient in set 2.
  - Then $\mathrm{GVIF}_1 = \det(\mathrm{R}_{b_1})\det(\mathrm{R}_{b_2})/\det(\mathrm{R}_b)$.
- We can apply the last result to other regression models, such as GLMs or the fixed effects in mixed models.
  - There is some slippage, however: The utopian situation is no longer uncorrelated xs (or sets of xs), as for linear least-squares, but otherwise similar data leading to uncorrelated *coefficients*.

Outline

1. Collinearity Diagnostics
2. Measuring Collinearity: Variance Inflation
3. Visualizing Collinearity
4. Generalized Variance Inflation
5. Dealing with Collinearity: No Quick Fix
6. References

Dealing with Collinearity: No Quick Fix

- Because collinearity is a problem with the data and not with the model, there generally isn't a satisfactory solution to the problem.
  - A regression model should be formulated to reflect hypotheses about the structure of the data or to put questions to the data.
  - If we include both x_1 and x_2 in a regression model, that should mean that we're interested in the partial relationship of y to x_1 holding x_2 constant, the partial relationship of y to x_2 holding x_1 constant, or both.
  - If x_1 and x_2 are so highly correlated ihat we can't adequately separate their effects, then there's little we can do short of collecting new data.
- There are nevertheless several strategies for dealing with collinear data.
  - Some of these approaches can be useful for pure prediction problems.
  - None of them magically make collinear data more informative.

Dealing with Collinearity: No Quick Fix

- *Model Respecification:* What I mean by "model respecification" in this context is removing xs from the regression equation to reduce collinearity, implicitly asking different questions of the data.
- *Variable Selection:* Variable-selection methods are automatic techniques for specifying a regression model by including only a subset of candidate explanatory variables.
  - When our interest is in *understanding* how the explanatory variables influence the response, rather than prediction, using a mechanical method to select the model automatically is not a reasonable strategy.
- *Regularization:* Regularization methods resolve the ambiguity produced by collinearity by driving (some) regression coefficients towards 0, with the goal of producing (biased) estimates with smaller MSE than the least-squares estimates.
  - The most common regularization methods in regression analysis are *ridge regression* (Hoerl and Kennard, 1970ab) and the *lasso* (Tibshirani, 1996).
  - If the goal isn't prediction, for regularization to achieve its goal, it's necessary to know something about the population regression coefficients, or implicitly to pretend to know.

Dealing with Collinearity: No Quick Fix

- *Prior Information About the βs*: Perhaps we're willing to make additional assumptions about the population regression coefficients.
  - These assumptions may take a very simple form, such as that two βs are equal, in which case we can get constrained least-squares estimates of the coefficients.
  - Or they may take the form of statements about what are a-priori plausible values for the βs, in which case we can employ Bayesian methods of estimation.
  - For these approaches to work, the prior information must be sufficiently specific to reduce the ambiguity due to collinearity, and we have to be honest about the state of our prior knowledge.
- These strategies have more in common than it might at first appear. For example:
  - Variable selection in effect respecifies the model, albeit mechanically.
  - Regularization (particularly the lasso) can drive coefficients to 0, effectively eliminating the corresponding regressors from the model.
  - Regularization also entails tacit assumptions about plausible values of the βs.

Outline

1. Collinearity Diagnostics
2. Measuring Collinearity: Variance Inflation
3. Visualizing Collinearity
4. Generalized Variance Inflation
5. Dealing with Collinearity: No Quick Fix
6. References

Collinearity Diagnostics

References

- J. Fox and G. Monette, *Generalized collinearity diagnostics*, Journal of the American Statistical Association **87** (1992), 178–183.

- A. E. Hoerl and R. W. Kennard, *Ridge regression: Biased estimation for nonorthogonal problems*, Technometrics **12** (1970a), 55–67.

- ______, *Ridge regression: Applications to nonorthogonal problems*, Technometrics **12** (1970b), 69–82.

- R. Tibshirani, *Regression shrinkage and selection via the lasso*, Journal of the Royal Statistical Society, Series B **58** (1996), 267–288.