If you must keep all your variables in the model, you should consider alternative methods of estimation such as ridge regression.

The effect of collinearity on prediction depends on where the prediction is to be made. The greater the distance is from the observed data, the more unstable the prediction. Distance needs to be considered in a Mahalanobis rather than a Euclidean sense.

Exercises

1. Using the `faithful` data, fit a regression of `duration` on `waiting`. Assuming that there was a measurement error in `waiting` of 30 seconds, use the SIMEX method to obtain a better estimate of the slope.
2. What would happen if the SIMEX method was applied to the response error variance rather than predictor measurement error variance?
3. Using the `divorce` data:
   (a) Fit a regression model with `divorce` as the response and `unemployed, femlab, marriage, birth` and `military` as predictors. Compute the condition numbers and interpret their meaning.
   (b) For the same model, compute the VIFs. Is there evidence that collinearity causes some predictors not to be significant? Explain.
   (c) Does the removal of insignificant predictors from the model reduce the collinearity? Investigate.
4. For the `longley` data, fit a model with `Employed` as the response and the other variables as predictors.
   (a) Compute and comment on the condition numbers.
   (b) Compute and comment on the correlations between the predictors.
   (c) Compute the variance inflation factors.
5. For the `prostate` data, fit a model with `lpsa` as the response and the other variables as predictors.
   (a) Compute and comment on the condition numbers.
   (b) Compute and comment on the correlations between the predictors.
   (c) Compute the variance inflation factors.

2. Using the `divorce` data, fit a regression model with `divorce` as the response and `unemployed`, `femlab`, `marriage`, `birth` and `military` as predictors.

   (a) Make two graphical checks for correlated errors. What do you conclude?

   (b) Allow for serial correlation with an AR(1) model for the errors. (Hint: Use maximum likelihood to estimate the parameters in the GLS fit by `gls(..., method="ML", ...)`. What is the estimated correlation and is it significant? Does the GLS model change which variables are found to be significant?

   (c) Speculate why there might be correlation in the errors.

3. For the `salmonella` dataset, fit a linear model with `colonies` as the response and `log(dose+1)` as the predictor. Check for lack of fit.

4. For the `cars` dataset, fit a linear model with `distance` as the response and `speed` as the predictor. Check for lack of fit.

5. Using the `stackloss` data, fit a model with `stack.loss` as the response and the other three variables as predictors using the following methods:

   (a) Least squares

   (b) Least absolute deviations

   (c) Huber method

   (d) Least trimmed squares

   Compare the results. Now use diagnostic methods to detect any outliers or influential points. Remove these points and then use least squares. Compare the results.