**Troubleshooting table**

|  |  |  |
| --- | --- | --- |
| **Problem** | **Diagnostics** | **Solutions** |
| Covariance matrix is nonpositive definite | * + Is there extreme collinearity among any of the variables?   + Are there outliers that inflate the correlations among variables?   + Are incomplete cases removed via pairwise deletion?   + Is there a transcription error?   + Is there a sampling error? | * + Try to correct any diagnostic red-flags from the previous column.   + *Advanced technique*: Perform ridge adjustment. |
| There is extreme collinearity. | * + Are composite scores being correlated with their constituent subscores?   + C:\D06B0CE5\483FF1FA-CFE3-4400-8A42-26C04D9FE4AF_files\image001.png   + Are any of these greater than 0.90?   + C:\D06B0CE5\483FF1FA-CFE3-4400-8A42-26C04D9FE4AF_files\image002.png   + Equivalently, is the variance inflation factor (VIF = 1/tolerance) > 10? | Since collinearity indicates that separate variables are measuring the same thing, try to eliminate the redundant variables or combine them into a single score. |
| There are outliers in the data | * + Do any observations fall more than 3 SD from the mean?   + Use a case analysis technique such as Mardia's index or Mahalanobis distance statistic in your statistics software. | * + *Check*: Are any of the outlying cases the result of a transcription or measurement error? Do the outlying cases actually belong to your sample's population?   + Remove the outliers   + Reduce the extreme value   + Transform the variable |
| The data does not follow a normal distribution | * + Does the data follow a normal distribution?   + Is the data skewed? (SI > 3)   + Is the data lepto- or platy-kurtic? (KI > 10) | Try a transformation of the data. |

**Data Screening exercise**

I have chosen to screen some reaction time data from a language processing task with 30–45 month-old children. I am interested in whether vocabulary size predicts reaction time. There are two trial conditions, one in which the child is prompted to look a familiar object named using a real word (e.g., *dog*) and another condition in which the child is prompted to look at unfamiliar object named with a nonsense word. Therefore the five variables in this dataset are two different measures of vocabulary, age in months, reaction time and trial condition.

These reaction times already underwent one iteration of screening and correction: Blinks and other random missingness in the eye-tracking data were imputed using neighboring data, RTs that were impossible fast (by virtue of how eye-movements work) were excluded and then RTs that were more than 2 SDs above the mean were dropped within each condition. Since reaction times are practically unbounded durations, trimming the slowest 5% of RTs seems appropriate.

Missingness was present in the dataset because not every trial yielded a usable reaction time, and a number of reaction times were trimmed as described above. It is possible that attention to the task predicts the number of usable reaction times, and therefore that the number of usable observations within each subject is not ignorable missing data. I checked against this possibility by regressing the number of reaction times onto condition, two measures of vocabulary, and age. There was a significant effect of age such that increasing one-month in age predicted an increase in usable data by 0.76 trials, controlling all other predictors. In other words, older kids may be disproportionately represented in the unaggregated data-set.

*Specific checks*

For the purposes of this exercise, I aggregated observations by condition within each subject by computing mean reaction times. (A more robust analysis would use these repeated measures to its advantage, of course.) The aggregated data contains no missing observations.

Collinearity was assessed by computing the squared multiple correlations of each variable as well as examining the bivariate correlations. All of these correlations were less than 0.90. The highest correlation was between expressive and receptive vocabulary measures, *r* = 0.73, which suggests that these scores measure different aspects of the same underlying vocabulary construct.

The covariance matrix showed positive eigenvalues, so it is positive definite.

Univariate normality was using measures for skew and kurtosis. Adequate values were found for these measures. Linearity and heteroscedacity were assessed for the additive model that regresses reaction time onto the four predictor variables, using a function that checks whether the GLM's assumptions hold for a model. All assumptions, including linearity and heteroscedacity, held for the additive model with log-transformed reaction times.

**Exercise 5**

The data are positively skewed, so greater values show heteroscedascity.

dataset <- list(Score = c(10:17, 27), Counts = c(6, 15, 19, 18, 5, 5, 4, 1,

1))

full\_data <- unlist(Map(rep, dataset$Score, dataset$Counts))

m <- lm(full\_data ~ 1)

plot.lm(m, which = 2)

