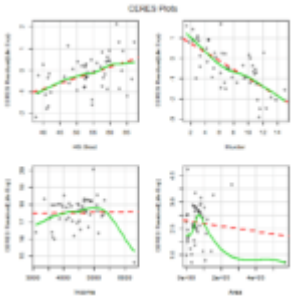
Multiple Regression (Part 3) Diagnostics



In the exercises below we cover some more material on multiple regression diagnostics in R. This includes added variable (partial-regression) plots, component+residual (partial-residual) plots, CERES plots, VIF values, tests for heteroscedasticity (nonconstant variance), tests for Normality, and a test for autocorrelation of residuals. These are perhaps not as common as what we have seen in Multiple Regression (Part 2), but their aid in investigating our model's assumptions is valuable.

Answers to the exercises are available here.

If you obtained a different (correct) answer than those listed on the solutions page, please feel free to post your answer as a comment on that page.

Multiple Regression (Part 2) Diagnostics can be found here.

As usual, we will be using the dataset state.x77, which is part of the state datasets available in R. (Additional information about the dataset can be obtained by running help(state.x77).)

First, please run the following code to obtain and format the data as usual:

```
data(state)
state77 <- as.data.frame(state.x77)
```

```
names(state77)[4] <- "Life.Exp"
names(state77)[6] <- "HS.Grad"
```

Exercise 1

For the model with Life.Exp as dependent variable, and HS.Grad and Murder as predictors, suppose we would like to study the marginal effect of each predictor variable, given that the other predictor is in the model.

a. Use a function from the car package to obtain added-variable (partial regression) plots for this purpose.

b. Re-create the added-variable plots from part a., labeling the two most influential points in the plots (according to Mahalanobis distance).

Learn more about multiple linear regression in the online course Linear regression in R for Data Scientists. In this course you will learn how to:

- Model basic and complex real world problem using linear regression
- Understand when models are performing poorly and correct it
- Design complex models for hierarchical data
- And much more

Exercise 2

a. Illiteracy is highly correlated with both HS.Grad and Murder. To illustrate problems that occur when multicollinearity exists, suppose we would like to study the marginal effect of Illiteracy (only), given that HS.Grad and Murder are in the model. Use a function from the car package to get the relevant added-variable plot.

b. From the correlation matrix in the previous Exercise Set, we know that Population and Area are the least strongly correlated variables with Life.Exp. Create added-variable plots for each of these two variables, given that all other

six variables are in the model.

Exercise 3
Consider the model with HS.Grad, Murder, Income, and Area as predictors. Create component+residual (partial-residual) plots for this model.

Exercise 4
Create CERES plots for the model in Exercise 3.

Exercise 5
As an illustration of high collinearities, compute VIF (Variance Inflation Factor) values for a model with Life.Exp as the response, that includes all the variables as predictors. Which variables seem to be causing the most problems?

Exercise 6
Using a function from the package lmtest, conduct a Breusch-Pagan test for heteroscedasticity (non-constant variance) for the model in Exercise 1.

Exercise 7
Re-do the test in the previous exercise by using a function from the car package.

Exercise 8
The test in Exercise 6 (and 7) is for linear forms of heteroscedasticity. To test for nonlinear heteroscedasticity (e.g., “bowtie-shape” in a residual plot), conduct White’s test.

Exercise 9
a. Conduct the Kolmogorov-Smirnov normality test for the residuals from the model in Exercise 1.
b. Now conduct the Shapiro-Wilk normality test.
Note: More Normality tests can be found in the nortest package.

Exercise 10

For illustration purposes only, conduct the Durbin-Watson test for autocorrelation in residuals. (NOTE: This test is ONLY appropriate when the response variable is a time series, or somehow time-related (e.g., ordered by data collection time.))