Regression Diagnostics Exercises

John Fox

FIOCRUZ November 2009

1. The data set `Moore` in the **car** package (see `?Moore`) contains data from a social-psychological experiment reported by Moore and Krupat (1971). The experiment was designed to determine how the relationship between conformity and social status is influenced by "authoritarianism." The subjects in the experiment were asked to make perceptual judgments of stimuli that were intrinsically ambiguous. Upon forming an initial judgment, the subjects were presented with the judgment of another individual (their "partner") who was ostensibly participating in the experiment; the subjects were then asked for a final judgment. In fact, the partner's judgments were manipulated by the experimenters so that subjects were faced with nearly continuous disagreement. The measure of conformity employed in the study (`conformity` in the data set) was the number of times in 40 critical trials that subjects altered their judgments in response to disagreement. The 45 university-student subjects in the study were randomly assigned to two experimental conditions (`partner.status`): In one condition, the partner was described as of relatively high social status (a "physician"); in the other condition, the partner was described as of relatively low status (a "postal clerk"). A standard authoritarianism scale (the "F-scale", `fscore`) was administered to the subjects after the experiment was completed. The authors divided the authoritarianism scores into three categories — low, medium, and high (`fcategory`).

(a) Analyze the data as the authors did, regressing `conformity` on dummy regressors for `fcategory` and `partner.status` (and their interaction) in a two-way ANOVA; look for unusual data in the two-way ANOVA.

(b) Analyze the data as a dummy regression (or analysis of covariance), regressing `conformity` on `fscore` and a dummy regressor for `partner.status` (and their interaction); look for unusual data in the dummy regression.

(c) Does it make sense to construct added-variable plots for the dummy regressors and interaction regressors in an ANOVA or dummy-regression model?

2. The data set `Prestige` in the **car** package (see `?Prestige`) contains data on occupational `prestige`, `income`, `education`, and percent `women` for 102 Canadian occupations around 1970. Perform a least-squares regression of `prestige` on the three other variables.

(a) Check for non-normality, non-constant error variance, and nonlinearity in this regression.Attempt to correct any problems that are detected. You might find the following strategy helpful: Use relatively simple diagnostics to check for problems and more sophisticated methods to follow up. To check for non-normality, construct a quantile-comparison plot and a kernel density estimate or histogram of the studentized residuals; to check for non-constant error variance, plot studentized residuals against fitted values; to check for nonlinearity, examine component+residual plots.

(b) Consider the following alternative analysis of the Canadian occupational prestige data: Regress `prestige` on `income`, `education`, percent `women`, and on dummy regressors for `type` of occupation (professional and managerial, white collar, blue collar); include interactions between `type` of occupation and each of `income`, `education`, and percent `women`. Why is it that the interaction between `income` and `type` of occupation can induce a nonlinear relationship between `prestige` and `income` when the interaction is ignored? (*Hint*: Construct a scatterplot of `prestige` vs. `income`, labeling the points in the plot by occupational `type`, and plotting the separate regression line for each occupational `type`.)

3. The data set `Chile` in the **car** package (see `?Chile`) contains data on a poll of voters conducted before the 1988 Chilean plebiscite, which in the event restored democracy to the country. Consider only voters intending to `vote` yes (`Y`) or no (`N`), recoding other voting intentions to missing. A yes vote represented support for the continuation of the then-current military government. You can use the `recode` command in the **car** package:

```
Chile$yes <- recode(Chile$vote, ' "Y" = "yes"; "N" = "no"; else = NA ')
```

or, equivalently,

```
Chile$yes <- with(Chile, factor(ifelse(vote == "Y", "yes",
              ifelse(vote == "N", "no", NA))))
```

Then regress `yes` on `region`, `population`, `sex`, `age`, `education`, and `income`, employing an additive logistic-regression model. (Do not use support for the status quo as an explanatory variable.)

(a) Use the diagnostic methods for generalized linear models described in the lecture to check the adequacy of the logistic-regression model that you fit to the data.

(b) The explanatory variables `income` and `population`, though numeric variables, are discrete (see, `table(Chile$income)` and `table(Chile$population)`). Test the linearity on the logit scale of the partial relationship between voting `yes` and each of these explanatory variables.