CPH 636 — Spring 2009 — Dr. Charnigo

Written Assignment 3 Solutions

Note: Your answers may differ from mine, depending on which observations you removed. I removed the observations with ID = 317, ID = 142, ID = 61, ID = 127, ID = 214, ID = 321, and ID = 394.

1a. None of the VIF plots for continuous explanatory variables exhibited appreciable compression of "E" points relative to "R" points (as defined in Lecture 5). Moreover, the largest VIF was 1.90, well below the threshold of 10 at which we begin to worry about collinearity. This is somewhat remarkable in that there were six explanatory variables related to heart rhythm: one would have anticipated some larger VIFs.

1b. The assumption of normality is questionable since the normal probability plot of studentized residuals exhibits a pattern characteristic of large kurtosis; however, I have seen much worse. The assumption of constant error variance is questionable since the plot of residuals against predicted values exhibits greater vertical spread for medium predicted values than for small predicted values, although the fact that there are relatively few small predicted values may lead to an exaggerated impression of heteroscedasticity; again, I have seen much worse.

1c to 1e. The simplest model for which C(p) is less than the number of parameters is the five-predictor model with explanatory variables Female, Height, QRSDur, QTInt, and PInt. I refer to this as Model I. The model for which C(p) is smallest, and for which the AIC is smallest, is the six-predictor model with explanatory variables Female, Height, QRSDur, QTInt, PInt, and ArrhythmiaAny. I refer to this as Model II. The model for which the SBC is smallest is the three-predictor model with Female, QRSDur, and QTint. I refer to this as Model III. The fitted models (in SAS syntax) are as follows:

```
PREDHRI = 177.96964 + 4.21905*Female + 0.16301*QRSDur - 0.26910*QTINT
- 0.16373*HEIGHT + 0.05763*PINT;
PREDHRII = 176.39843 + 4.54485*Female + 0.14264*QRSDur - 0.26748*QTINT
- 0.15746*HEIGHT + 0.06431*PINT + 2.09667*ARRHYMTHIAANY;
PREDHRIII = 153.21276 + 5.57060*Female + 0.17821*QRSDur - 0.26619*QTINT;
```

1f. I felt that the augmented partial residual plots for QTInt were suggestive of a quadratic term. I say "plots" rather than "plot" because I created new augmented partial residual plots when fitting Models I, II, and III. However, looking at the original augmented partial residual plot from Written Assignment 2 would have been acceptable. [Disadvantage of looking at the original plot: deletion of some explanatory variables may affect our impression about whether a quadratic term in another explanatory variable is necessary. Advantage of looking at the original plot: one may see that a deleted explanatory variable was removed because the linear approximation to its relationship with the response variable was completely inadequate, in which case one may consider reinstating the deleted explanatory variable and including a quadratic term in it.]

The augmented partial residual plots for Height and QRSDur were questionable, but I was not convinced to add a quadratic term since the appearance of a nonlinear relationship was largely created by one point.

The augmented partial residual plots for PInt were not suggestive of a quadratic term.

Obviously we would not consider including a quadratic term in a dichotomous explanatory variable such as Female or ArrhythmiaAny.

Hence, I defined Models IV, V, and VI to be like Models I, II, and III except that I added a quadratic term in QTInt. The fitted models (in SAS syntax) are as follows:

```
PREDHRIV = 448.85104 + 3.59025*Female + 0.10492*QRSDur - 1.73146*QTINT
- 0.14309*HEIGHT + 0.07228*PINT + 0.00197*QTINT2;
PREDHRV = 442.58491 + 3.74072*Female + 0.09752*QRSDur - 1.70053*QTINT
- 0.14088*HEIGHT + 0.07479*PINT + 0.88461*ARRHYMTHIAANY + 0.00192*QTINT2;
PREDHRVI = 420.10952 + 4.69526*Female + 0.11802*QRSDur - 1.68166*QTINT + 0.00190*QTINT2;
```

2. For the validation data set I had $\sum_{i=1}^{n}(y_i - \bar{y})^2 = \sum_{i=1}^{n}(y_i - 75.0982143)^2 = 21949.92$ and $\sum_{i=1}^{n}(y_i - \hat{y}_i)^2 = 10417.60$ for Model I, $= 10192.43$ for Model II, $= 11285.90$ for Model III, $= 10116.48$ for Model IV, $= 10152.19$ for Model V, and $= 11248.66$ for Model VI. Hence, I had $R^2_{valid} = 0.5253923$ for Model I, $= 0.5356507$ for Model II, $= 0.4858341$ for Model III, $= 0.5391108$ for Model IV, $= 0.537484$ for Model V, and $= 0.4875307$ for Model VI. I chose Model IV.

3. For the test data set I had $\sum_{i=1}^{n}(y_i - \bar{y})^2 = \sum_{i=1}^{n}(y_i - 72.9285714)^2 = 12391.43$ and $\sum_{i=1}^{n}(y_i - \hat{y}_i)^2 = 12115.26$ for Model IV. Hence, I had $R^2_{test} = 0.02228718$ for Model IV.

An obvious question: Why did Model IV look pretty good on the training and validation data sets but awful on the test data set? The answer is that there was a single observation in the test data set for which the squared prediction error was 4594.59. This was the subject with ID 100, who had the lowest QTInt (241) in the test data set and thus an extremely high (in fact, unreasonably high: 139.783) predicted value for HeartRate. If this one subject had not been in the test data set, the R^2_{test} would have been a far more palatable 0.3930327.

Lessons to be learned? (a) There is some risk of being harmed by a quadratic term, especially in an explanatory variable with high kurtosis. (b) Part of the problem is that precisely estimating parameters in models with quadratic terms is difficult, as exemplified by training data VIFs of 205.92398 and 207.94274 for QTInt and its quadratic term. (c) Perhaps the metric we are using to quantify predictive ability (the sum of squared prediction errors or the R^2 derived from it) is too sensitive to a single bad prediction. (d) Perhaps linear regression is not the best approach for making predictions when the explanatory variables have high kurtosis.

4a to 4c. The model for which the AIC was smallest was the six-predictor model with explanatory variables QRSDur, PInt, TInt, Weight, Female, and HeartRate. I refer to this as Model I. Model I provided 70.5% correct classification at a threshold of 0.65 and 90.4% sensitivity, 17.2% specificity at a threshold of 0.20. The model for which the SC was smallest was the two-predictor model with explanatory variables QRSDur and PInt. I refer to this as Model II. The fitted models (in SAS syntax) are as follows:

```
ESTLOGITRISKI = -5.1602+0.0596*QRSDUR-0.0177*PINT+0.0108*TINT
-0.0225*WEIGHT-0.6878*FEMALE+0.0191*HEARTRATE;
ESTLOGITRISKII = -4.8684+0.0717*QRSDUR-0.0178*PINT;
```

4d. The LOGISTIC macro did not provide me with any logit and/or partial delta logit plots, so I chose Model III to be a sensible competitor based on reasonably low values for both the AIC and the SC, namely the four-predictor model with explanatory variables QRSDur, PInt, TInt, and Weight. The fitted model (in SAS syntax) is as follows:

```
ESTLOGITRISKIII = -5.6197+0.0697*QRSDUR-0.0160*PINT+0.0118*TINT-0.0181*WEIGHT;
```

5. Model I provided 72.32% correct classification on the validation data at a threshold of 0.50, Model II provided 67.86% correct classification at thresholds between 0.45 and 0.55, and Model III provided 69.64% correct classification at thresholds between 0.45 and 0.55. I chose Model I.

6. Model I provided 70.54% correct classification on the test data at a threshold of 0.55. Also, we had 93.62% sensitivity and 3.08% specificity at a threshold of 0.15. The latter findings are very disappointing: at a threshold of 0.15, the fraction of people in the test data predicted to have arrhythmia was higher among those who did not actually have it than among those who did!