CPH 636 — Spring 2009 — Dr. Charnigo

Written Assignment 2 Solutions

1a to 1c. My answers are below. Impressions about the shape of a distribution are somewhat subjective. Observations with ID = 317 and ID = 142 contain gross mistakes (608 and 780 on Height) and should definitely be removed prior to exericse 2. Observations with ID = 61, ID = 127, ID = 214, ID = 321, and ID = 394 are also highly suspicious (Height seems too large given Age and Weight, Weight seems too small given Height and Age, Weight seems too large absolutely, Weight seems too small given Height and Age, PRInt seems too large absolutely). I removed them but would not fault you for leaving them in.

Variable	Mean	SD	Shape	Outliers
Age	46.47	16.47	near normal	317, 61, 142
Height	166.19	37.17	contaminated by gross mistakes	5, 347, 298, 431, 1, 433, 402, 127, 430, 449, 317, 211, 424, 380, 425, 114, 321, 61, 333, 404, 199, 142
Weight	68.17	16.59	high kurtosis	402, 127, 430, 317, 214, 255, 211, 309, 380, 425, 321, 61, 404, 142
QRSDur	88.92	15.36	strong positive skew and very high kurtosis	298, 431, 86, 301, 208, 313, 371, 427, 380, 425, 389, 190, 90, 294, 3, 89, 404, 396, 428
PRInt	155.15	44.84	very high kurtosis	272, 351, 111, 354, 299, 301, 72, 285, 421, 309, 183, 394, 96, 107, 134, 65, 49, 21, 311, 429, 109, 220, 8, 218, 244, 57, 175
QTInt	367.21	33.39	high kurtosis	242, 303, 361, 217, 208, 317, 313, 371, 434, 421, 309, 99, 183, 425, 180, 389, 422, 321, 360, 382, 390, 61, 404, 142, 100, 396
TInt	169.95	35.63	positive skew and high kurtosis	375, 433, 260, 86, 30, 352, 205, 208, 371, 226, 330, 324, 368, 425, 253, 247, 411, 389, 190, 295, 95, 90, 29, 336, 297, 369, 168, 381, 357, 89, 186, 373, 281, 236, 258, 152, 396, 362, 428
PInt	90.00	25.83	high kurtosis	272, 111, 303, 354, 163, 299, 301, 352, 72, 421, 309, 394, 96, 107, 103, 21, 311, 178, 418, 356, 281, 109, 220, 2, 199, 286, 218, 244, 57, 175, 4
HeartRate	74.46	13.87	positive skew and high kurtosis	317, 434, 380, 349, 321, 382, 404, 388, 142, 244

1d. There are 114 observations with outlying values on at least one of the variables. Observations with ID = 142, ID = 317, ID = 404, and ID = 425 have outlying values on five of the variables! If there were 90 variables instead of nine, many more observations would have an outlying value on at least one variable.

To get some perspective on this, suppose that $Z_1, Z_2, \ldots, Z_{90}$ are independent and standard normal.

Taking ± 2.698 as a theoretical cutoff for an outlier — because $z_{0.25} - 1.5(z_{0.75} - z_{0.25}) = -2.698$ and $z_{0.75} + 1.5(z_{0.75} - z_{0.25}) = 2.698$ — yields the following computation for the probability that at least one of $Z_1, Z_2, \ldots, Z_{90}$ attains an outlying value: $1 - P(|Z_1|, |Z_2|, \ldots, |Z_{90}| \leq 2.698) = 1 - P(|Z_1| \leq 2.698)^{90} = 1 - 0.9930^{90} = 0.4675$. Most of the existing variables in this data set have greater than normal kurtosis. Hence, the 0.4675 is a very conservative estimate of the fraction of observations that would have an outlying value on at least one variable if there were 90 variables and these 90 variables exhibited similar levels of kurtosis.

What I have described above reflects the curse of dimensionality.

2. Out of 452 observations, there are 160 (35.40%) for females without arrhythmia, 89 (19.69%) for females with arrhythmia, 85 (18.81%) for males without arrhythmia, and 118 (26.11%) for males with arrhythmia.

I accepted any graphical output from the FREQ macro.

3a. Your numerical answers may differ slightly from mine, depending on which observations you removed. As previously stated, I removed the observations with ID = 317, ID = 142, ID = 61, ID = 127, ID = 214, ID = 321, and ID = 394. Five of these were in the training subset, one was in the validation subset, and one was in the test subset. The following computations apply to the training subset.

Only one of the explanatory variables has even a moderately strong (greater than 0.20 in absolute value) correlation with HeartRate. That is QTInt, for which the correlation is -0.5987. The negative correlation seems to make sense: if part of a cycle in the heart rhythm is longer, then we anticipate fewer beats per minute.

We are most comfortable reporting correlations for cardinal variables, especially those whose distributions are near normal. Thus, we may be uneasy reporting correlations involving the dichotomous variables Female and ArrhythmiaAny.

3b. Again, your numerical answers may differ slightly from mine.

We have $R^2 = 0.4573$, meaning that 45.73% of the variation in HeartRate (within the training subset) is accounted for by the multiple linear regression model with 10 explanatory variables. We have adjusted $R^2 = 0.4313$ and predicted $R^2 = 1 - 22717/36611 = 0.3795$. Predicted R^2 is smaller than ordinary R^2, which is unsurprising because predicted R^2 is intended to reduce the optimism of ordinary R^2.

3c. Small and large predictions have a greater tendency to err on the low side (more blue dots above the red line in the far left and far right portions of the graphic), while medium predictions have a greater tendency to err on the high side (more blue dots below the red line in the middle portion of the graphic).

3d. In what follows, symbols like y_i refer to HeartRate.

The residuals $y_i - \hat{y}_i$ (shown in blue) and the residuals $y_i - \hat{y}_{i(i)}$ based on predictions from auxiliary multiple linear regression models from which individual observations are excluded (shown in red) are usually of the same sign as the centered response values $y_i - \bar{y}$ (shown in black) and are typically of somewhat lesser magnitude. Thus, the dark part of the graphic is mostly but not completely obscured by the light part, indicating that the predictions $\hat{y}_i$ or $\hat{y}_{i(i)}$ are modestly better than $\bar{y}$. In other words, the multiple linear regression model is of some value for making predictions, although much of the variation in HeartRate remains unexplained by the multiple linear regression model.

3e. Even though ArrhythmiaType has 16 possible values, they are not arranged in a medically meaningful order. Normal ("1") is better than Ventricular Premature Contraction ("7"), but Coronary Artery Disease

("2") is worse than Ventricular Premature Contraction ("7"). So, without some careful effort in reassigning the designations "1" through "16", we must regard ArrhythmiaType as a nominal variable — not an ordinal variable.

The problem with using a nominal variable as an explanatory variable in linear regression is illustrated as follows. Suppose, for concreteness, that we have postulated $E[Y|X = x] = \alpha + \beta x$ with X equal to ArrhythmiaType and Y equal to HeartRate. If $\beta \geq 0$, then $E[Y|X = 1] \leq E[Y|X = 2] \leq E[Y|X = 7]$. If $\beta \leq 0$, then $E[Y|X = 1] \geq E[Y|X = 2] \geq E[Y|X = 7]$. In any case, there is no allowance for the possibility that $E[Y|X = 2]$ may be greater than both $E[Y|X = 1]$ and $E[Y|X = 7]$. A statistical model that restricts $E[Y|X = 2]$ to be between $E[Y|X = 1]$ and $E[Y|X = 7]$, when there is no medical rationale for such a restriction, is defective.

Of course, no such problem arises if the nominal variable has only two categories, and this leads us to a solution. Let $Z_j := 1$ if $X = j$ and $Z_j := 0$ otherwise, for $j \in \{1, 2, \ldots, 15\}$. Then use $Z_1, Z_2, \ldots, Z_{15}$ as explanatory variables instead of X. Now there is no restriction on $E[Y|X = 1]$, $E[Y|X = 2]$, and $E[Y|X = 7]$. The difference between the first two quantities is $\beta_1 - \beta_2$, while the difference between the last two quantities is $\beta_2 - \beta_7$. Just because $\beta_1 - \beta_2$ is positive does not mean that $\beta_2 - \beta_7$ must be positive, or vice versa.