STA 302 / 1001 (A. Gibbs)
Solutions to Exercises in Chapter 6 of Sheather

1. $\text{Var}(\hat{\mathbf{Y}}|\mathbf{X}) = \text{Var}(\mathbf{HY}|\mathbf{X}) = \mathbf{H}\text{Var}(\mathbf{Y}|\mathbf{X})\mathbf{H}' = \mathbf{H}\sigma^2\mathbf{I}\mathbf{H} = \sigma^2\mathbf{H}$

2. Many of the predictor variables are related to each other, so multicollinearity will be a problem and it will be difficult to sort out which variables are important and to interpret the coefficients. Although year is added to allow for inflation, there is likely also correlation over time between the observations on the same team. So the assumption of uncorrelated error terms is likely violated.

3. A couple of notes about this exercise:
   (1) Since the textbook computer output didn't included partial regression plots, I decided also not to provide them; that is not meant to undermine their usefulness.
   (2) Both matrices of scatterplots in the textbook are missing the response variable. I don't know why.
   (3) My output from SAS for the transformed data does not agree with the output given in the textbook. I can't figure out why. I think there is a problem in the textbook.
   (4) The question says that Model (6.37) was decided on from the Box-Cox method. We can still evaluate the models without worrying about this.

   (a) It is clear that this model is not appropriate from both the scatterplots of the dependent versus the independent variables and from the plots of the standardized residuals versus the independent variables. Many of these plots show evidence of curvature and all show non-constant variance.

   (b) More important than any possible curvature is the increasing variance in this plot indicating that a transformation of the response variable should be considered.

   (c) Observations 66, 67, 88, and 91 all have high leverage (h_{ii} well above the cut-off of $2(p+1)/n = 0.068$). They are "bad" leverage points if they are also influential. This is true of observation 67 which has a Cook's distance of 0.223, well above all the values of Cook's distance for the other points. (Observation 67 is the Honda Insight.)

   (d) The pairwise scatterplots are much improved for the transformed data, showing no curvature. There are a few unusual points. The unusual points are evident in the residual plots. In particular, there are two points with very large standardized residuals (close to 4). These are observations 67 and 222 (the Mercedes-Benz CL500). Observation 67 is still influential (large Cook's distance). Except for these points, the normal quantile plot looks very good. It may be worth investigating why these 2 cars are so unusual, and whether it makes sense to treat them differently, possibly removing them from the model.

(e) The test statistic is $\frac{(7.23583-7.23365)/2}{0.03201} = 0.03405$. Under the null hypothesis that the coefficients of the two omitted terms are 0, this is an observation from an $F(2, 226)$ distribution. The p-value is very large (no need to calculate it!), so the data are consistent with coefficients of 0 for these two terms. So it is reasonable to remove them from the model. cannot reject

(f) Indicator variables would be needed to be added for the manufacturers. One less indicator variable than the number of manufacturers is needed.

4. (a) Looking at scatterplots of each predictor variable with the response (Krafft point), there are two concerns about the validity of a linear model: (1) There are two points (with values of Krafft in the middle, large RA, and small VTINV, DIPINV, and HEAT) which do not follow the pattern of the other points. They may be influential, or may be outliers as they may not follow the pattern of the fitted regression. (2) The points appear to come in two groups. They may be better captured by including an indicator variable for the grouping in the equation. It is certainly worth investigating whether there is a physical reason why they group like that.

Looking at the plots of the standardized residuals versus predicted values and versus the explanatory variables, the only apparent problem is the two points on the left in the plot of the residuals versus HEAT. They may be influential. There are no outliers (residuals are at most 2 in absolute value) and there is no evidence of curvature or non-constant variance.

The normal quantile plot shows some departures from normality in the tails.

The added variable plots show linear relationships, although the relationship with VTINV is weak. They do not show curvature. The points in the far right of the DIPINV plot may be influential.

Looking at the influence statistics, there are a few points that have high leverage. And as noted in plots above there are indeed influential points, particularly observations 12 and 32. Observation 32 is one of the points noted in the scatterplots, while observation 12 has large values of VTINV, DIPINV and HEAT and small value of RA.

There is serious multicollinearity as one can see in the scatterplots which show strong relationships among the explanatory variables and with the large values of the variance inflation factors

(b) Despite what the question says, there is no curvature in the residual plots (unless one considers seriously overfitting the data). If there was curvature, it would be appropriate to transform the explanatory variable, or add higher order terms in the explanatory variable to the model.

(c) The criteria suggested are naive. Assuming that r is the square root of R^2, it increases with the number of predictors. None of the suggested criteria give

any consideration to whether or not the linear model is appropriate (curvature, outliers, influential points). The values of the independent variables are not distributed such that there are some unusually large, potentially influential observations.

5. (a) Transforming Y (prize money) and not any of the independent variables seems like a reasonable place to start. The behaviour of Y will be improved by a transformation that adjusts the spacing by bringing in the right tail of the distribution. As well, it may improve the indications of increasing variance.

   (b) Using the log of prize money as the response, the scatterplots show that it seems reasonable to consider linear relationships with each of the predictors. There are no obvious indications of influential points or outliers.

   The plots of the residuals versus the predicted values and explanatory variables do not show any problems with curvature or non-constant variance. There is at least one point with a fairly large standardized residual (> 3) and one observation with an unusually high predicted value. The normal quantile plot shows no departures from normality.

   The added variable plots indicate that linear models are appropriate, but that some of the variables are not important, given that the others are in the model (in particular, driving accuracy, putting average, and number of putts per round).

   (c) The large residual is for observation 185. This observation also appears to be influential (large Cook's distance). (This is Tom Lehman, who made more prize money than his statistics indicate he should. It's not Tiger Woods!)

   (d) In addition to the unusual point mentioned above, there is evidence of multicollinearity with some high variance inflation factors. Having correlated predictors in the model reduces the efficiency of the estimates (by increasing their standard error) and makes it difficult to understand the effects of the individual explanatory variables on the response.

   (e) Based on the t-tests for the coefficients, only one predictor at a time should be removed from the model. This is because each of the t-test statistics is calculated with all other variables in the model and the test considers the effect of the variable over and above the other variables in the model. Removing one variable may result in a variable which had a large p-value now having a much smaller p-value. Since we have multicollinearity, this is particularly likely to be the case. Multiple variables can be removed from the model in one step only after a partial F-test indicates that the data are consistent with all coefficients tested being 0.