STA 302 / 1001 (A. Gibbs)
Sketch of Solutions to Exercises in Chapter 3 of Sheather

1. (a) While a straight line model appears to fit the data well, the residual plot reveals patterns in the data that the straight line model does not adequately describe. There is one point that is unusual and not fit well by the model, and for the rest of the points there is a pattern to where the fitted model systematically over- and under-estimates the observed data.

   The influence statistics also indicate some problems with the data. No points have overly large values of h_{ii} so this does not flag points with high leverage ($4/n = 0.235$ and the largest values of h_{ii} are 0.237 and 0.243 which are close to the cutoff). However, the $13th$ and $17th$ points are influential. For DFFITS and DFBETAS, the cut-off is 1 since this is a small dataset. For the $13th$ observation, DFFITS is 2.43 and DFBETAS for the slope is 2.12. For the $17th$ observation, DFFITS is -1.26 and DFBETAS for the slope is -1.10. Cook's distance is also large for these 2 observations; its is 1.37 for the $13th$ observation and 0.630 for the $17th$ observations (cut-off is 0.235). The standardized residuals are large for these two observations (4.39 for the $13th$ observations and -2.27 for the $17th$ observation) indicating that the fitted model does not fit these points well.

   While the influence statistics indicate problems, the nature of the problems is identified in the first paragraph. The $13th$ observation is the unusual observation identified in the plot of the standardized residuals versus the predicted values and the large influence statistics for the $17th$ observation are a reflection of the systematic pattern in the residuals.

   (b) The point with the large positive residual could be removed or dealt with separately (after some investigation for why it is unusual) as it does not fit the pattern of the rest of the data. The rest of the data should be modeled with a curvilinear relationship such as a quadratic function.

2. Because of increasing variance, a transformation of Y, such as square root or log, is appropriate. This may also fix the curvilinear pattern in the data. However, if the relationship between y and x is not monotone, it is necessary to fit a quadratic model (in x) with the transformed y.

3.

4. **First model**

   (a) No. Looking at the influence statistics, the 3rd and 31st observations are influential with large values of leverage, Cook's distance, DFFITS, and DFBETAS for the slope. (This dataset is fairly small, so I'd use the small dataset cut-offs, but a conservative approach may also look at the cut-offs for larger datasets.) These

two influential points have relatively large values of the explanatory variable, and can be seen in the scatterplot and plot of the standardized residuals versus the explanatory variable. However, I am not going to pay too much attention to these points because there are other indications of problems. The normal quantile plot indicates heavy tails in the distribution of the residuals. Of most concern is that there is increasing variance seen in the plot of the standardized residuals versus the explanatory variable, and in the plot of the square root of the absolute value of the standardized residuals versus the explanatory variable which has an increasing pattern.

(b) The interval would be too short. The calculation of the prediction interval would have been made assuming the variance was the same for all values of tonnage. However the variance is increasing with tonnage. Since 10,000 is a relatively large value for tonnage, the estimate of the error variance under the assumption of constant variance is too small there. So the margin of the error of the prediction interval would be too small.

Second model

The second model is a definite improvement. None of the influence statistics are very large and even those that are large are much closer to the cut-offs. There is no longer any sign of increasing variance. From the normal quantile plot, the right tail of the distribution of the residuals is a little lighter than a normal distribution, but it is closer to normally distributed than the first model.

5. (a) No! The variance of the residuals increases with dealer cost so inferences will not be valid.

(b) While it is possible to identify other problems (there are influential points, points with very unusually large residuals, heavy tails in the distribution of the residuals), the key problem that needs to be fixed before these should be considered is the increasing variance.

(c) It is an improvement, particularly in the variance and the residuals are closer to normally distributed.

(d) Back-transforming the fitted equation gives

$$\text{suggested } \hat{\text{retail price}} = e^{-0.06946}(\text{dealer cost})^{1.01484}$$

Changing dealer cost by a factor of k changes the suggested retail price by a factor of $k^{1.01484}$ on average. So, for example, doubling dealer cost results on average in the suggested retail price increasing by a factor of 2.021.

(e) There are still influential points and the normal quantile plot indicates a group of points in the left tail with unusual residuals. These points (with large negative residuals) can also be seen in the plot of the standardized residuals versus the

explanatory variable. These are the points with the largest Cook's distance. They are the 4 Suzuki cars. The regression model overestimates the prize of these cars. Perhaps they should be treated differently in the model, or maybe other variables other than dealer cost should be included in the model that will help explain the prize of these cars.

6. Not covering.

7. Need to find the function $f(\mu)$ such that

$$f(\mu) \propto \int \frac{1}{\mu}\, d\mu$$

which is $\log(\mu)$.

8. Overall, the simple linear regression model without transformation fits the data fairly well with no obvious violations of the assumptions. There are some fairly large residuals, but they are not unusually large. None of the values for Cook's distance are over the cut-off for influential statistics. Some of the values of DFFITS and DFBETAS are over the large dataset cut-offs, but the dataset is not large (only 49 observations). While some of the leverage values are over the cut-off, there are no strikingly large values and, as noted, these points are not influential. The plots of standardized residuals do not show evidence of curvature or non-constant variance or unusually large residuals. The normal quantile plot does not indicate that there are deviations from normality.

   The paper referenced in the question expresses concern about the negative intercept (how can price be negative?) and tries to fix that. But as long as we are careful not to extrapolate, this model seems to fit the data fairly well and inferences will be valid.