STAT 212-501 (Fall 2022)

Homework 2

Problem 1. (Working through some details of a simple 'log-linear' model, i.e. a linear model for $\log Y$.)

In class, we discussed a lot about usage of $\log(\cdot)$ transformations and their possible benefits in regression analyses. We will work through one such exercise here under a very simple setting. Suppose we have a response variable $Y > 0$ and a single predictor variable x. Consider the following **'log-linear' model:**

$$\log Y_i = \beta_0 + \beta_1 x_i + \epsilon_i, \quad i = 1, \ldots, n, \quad \text{where } \epsilon_i \sim N(0, \sigma^2) \text{ and } \epsilon_1, \ldots, \epsilon_n \text{ are independent.}$$

(a) Show rigorously that for the model above, the least square estimates $\hat{\beta}_0$ and $\hat{\beta}_1$ are given by:

$$\hat{\beta}_1 = \frac{\sum_{i=1}^n (\log Y_i - \bar{Y}_{\log})(x_i - \bar{x})}{\sum_{i=1}^n (x_i - \bar{x})^2} \text{ and } \hat{\beta}_0 = \bar{Y}_{\log} - \hat{\beta}_1 \bar{x}, \quad \text{where } \bar{Y}_{\log} = \frac{1}{n} \sum_{i=1}^n \log Y_i.$$

Problem 2. Use the 'Baseball Salary Data' (with `salary` as the response) to do the following:

(a) Use the command `'leaps'` in the `R` package `''leaps''` along with the strategy discussed in class to choose a good subset of the 16 predictors to include in a linear model with `''salary''` as the response (and **not** its logarithm). **Describe fully the rationale you use** in choosing your model.

(**Note:** A help file on the `''leaps''` package, as well as the `R` function `'leaps.AIC'` (used to compute the AIC and BIC values for the best subset models of each size returned by `'leaps'`), are available under `Codes` on `Canvas`. Feel free to use these resources for part **(a)** above.)

(b) Plot the *standardized* residuals from your chosen model (in part **(a)** above) versus an index running from 1 to 337. Identify any players who have standardized residuals that are larger in absolute value than 3. Are these players different in any important way from most of the other players?

(c) Provide (i) a plot of the standardized residuals versus the predicted values and comment on the plot, and (ii) a Normal probability (or Q-Q) plot of the standardized residuals and comment on the plot.

(d) Provide a plot of Cook's D values. Do any data points seem to be influential? Why or why not?
