PSTAT 126: Homework #3

1) Obtaining data samples from the dataset “HW 3 Dataset 1” (which will be sent in an email with that subject line), utilize any or all of the various regression model-fit assessment visual plots and/or graphs, results of statistical hypothesis tests, and numerical regression model accuracy measures, as well as any other applicable diagnostic tools available in R that we have discussed in the course thus far, to help identify and validate a linear regression model that appears to best fit the given data. These will likely include for example direct X-Y variable graphs and residuals vs. fitted value plots implemented in R. You will likely need to try a number of different response and/or predictor variable transformations, for example, using the diagnostic tools to decide which regression models appear consistent with the data and which do not. You should identify one or perhaps two candidate regression models that you believe are most consistent with the given data. Please try to include screenshots of the graphical plots you use as well as quote any relevant R code output. You can or even should include the R code itself as well if you feel it helps to support your argument. Please explain your reasoning as to why the model(s) you propose may be the right one(s) in plain natural language writing.

2) We have been finding a set of optimal parameters for the linear least-squares regression minimization problem by identifying critical points, i.e. points at which the gradient of a function is the zero vector, of the following function:

$$F(\alpha_0, \dots, \alpha_M) = \sum_{n=1}^{N}(\alpha_0 + \alpha_1 x_{n1} + \cdots + \alpha_M x_{nM} - y_n)^2 .$$

Help to justify this methodology in the following way. Letting $G: \mathbb{R}^d \to \mathbb{R}$ be any function that is differentiable everywhere, show that, if G has a local minimum at a point x_0, then its gradient is the zero vector there, i.e., $\nabla G(x_0) = \mathbf{0}$.

3) A function $G: \mathbb{R}^d \to \mathbb{R}$ is called *convex* if it satisfies the following:

$$G(\lambda u + (1-\lambda)v) \leq \lambda G(u) + (1-\lambda)G(v),$$

for any $u, v \in \mathbb{R}^d$ and λ any scalar with $\lambda \in [0,1]$. Show that the function F in Question 2 above is convex using the following steps:

a) Show that the sum of convex functions is convex.

b) Define the function

$$H(\alpha_0, \dots, \alpha_M) = \alpha_0 + \alpha_1 x_1 + \cdots + \alpha_M x_M - y,$$

for any scalar, fixed constants $x_1, \dots, x_M, y$. Show that

$$H(\lambda u + (1-\lambda)v) = \lambda H(u) + (1-\lambda)H(v),$$

where u, v, and λ are as at the beginning of this question (Question 3) with $d = M + 1$.

c) Show that the function $f(t) = t^2$, t a scalar real number, is convex.

d) Explain how steps (a)-(c) can be put together to establish that the function F in Question 2 is convex.

4) Show that if a convex function $G: \mathbb{R}^d \to \mathbb{R}$ has a local minimum at a point x_0, then this local minimum is also a global minimum for the function over all of its domain $\mathbb{R}^d$.

5) Obtaining data samples from the dataset “HW 3 Dataset 2” (which will be sent in an email with that subject line), utilize any or all of the various regression model-fit assessment visual plots and/or graphs, results of statistical hypothesis tests, and numerical regression model accuracy measures, as well as any other applicable diagnostic tools available in R that we have discussed in the course thus far, to help identify and validate a linear regression model that appears to best fit the given data. These will likely include for example direct X-Y variable graphs and residuals vs. fitted value plots implemented in R. You will likely need to try a number of different response and/or predictor variable transformations, for example, using the diagnostic tools to decide which regression models appear consistent with the data and which do not. You should identify one or perhaps two candidate regression models that you believe are most consistent with the given data. Please try to include screenshots of the graphical plots you use as well as quote any relevant R code output. You can or even should include the R code itself as well if you feel it helps to support your argument. Please explain your reasoning as to why the model(s) you propose may be the right one(s) in plain natural language writing.