	Name: Hye-Young, Park **Class: STA502: Math & Stat for MBA** **Date: 2021/09/06** **Problem Set 1 - Question 4**
QUESTIONS	***NOTES***
Question 4. **1.** If all the observations on a particular explanatory variable are multiplied by λ, then the residuals of the regression are unchanged while the corresponding regression coefficient is multiplied by 1/λ. Use this result to explain what will happen when a particular explanatory variable is measured in thousands of kgs instead of millions of kgs.	$y = X\beta + \varepsilon$ can be represented in regression model as below. $y = \beta_1 x_1 + \beta_2 x_2 + \beta_3 x_3 + \cdots + \beta_k x_k + \varepsilon$ New explanatory variables which multiplied by λ are substitute x_i: $x_i = \lambda x'_i$ $y = \beta_1 \lambda x'_1 + \beta_2 \lambda x'_2 + \beta_3 \lambda x'_3 + \cdots + \beta_k \lambda x'_k + \varepsilon$ Then the new corresponding regression coefficients are: $\beta'_i = \lambda \beta_i$ $\beta_i = \frac{1}{\lambda} \beta'_i$ $y = \beta'_1 x'_1 + \beta'_2 x'_2 + \beta'_3 x'_3 + \cdots + \beta'_k x'_k + \varepsilon$ If all the observations on a particular explanatory variable are multiplied by λ, then the residuals of the regression are unchanged while the corresponding regression coefficient is multiplied by 1/λ. 1 million kg = 1000 thousand kg $x_i = 1000 x'_i$ Thus, the residuals of the regression are unchanged while the corresponding regression coefficient is multiplied by 1/1000.
2. If a constant λ is added to all observations of a particular explanatory variable in a regression containing a constant term, then the corresponding regression coefficient is unchanged. Is any other coefficient affected? Use this result to explain that the coefficient of an explanatory variable appearing in a regression in logarithmic form, the corresponding coefficient is independent of the units in which the variable is measured.	$y = X\beta + \varepsilon$ can be represented in regression model as below. $y = \beta_0 + \beta_1 x_1 + \beta_2 x_2 + \beta_3 x_3 + \cdots + \beta_k x_k + \varepsilon$ New explanatory variables which multiplied by λ are substitute x_i: $\ln A_i = x_i = x'_i + \lambda$ $y = \beta_0 + \beta_1 (x'_1 + \lambda) + \beta_2 (x'_2 + \lambda) + \cdots + \beta_k (x'_k + \lambda) + \varepsilon$ $= \beta_0 + \lambda \sum_{i=1}^{k} \beta_i + \beta_1 x'_1 + \beta_2 x'_2 + \beta_3 x'_3 + \cdots + \beta_k x'_k + \varepsilon$ Then the new corresponding regression coefficients are: $\beta'_i = \beta_i$ The corresponding regression coefficient is unchanged. But, product of λ and the summation of regression coefficient is added to the constant term. $\beta'_0 = \beta_0 + \lambda \sum_{i=1}^{k} \beta_i$ $y = \beta'_0 + \beta_1 x'_1 + \beta_2 x'_2 + \beta_3 x'_3 + \cdots + \beta_k x'_k + \varepsilon$ $e^y = B = e^{\beta_0} A^{\beta_1} A^{\beta_2} \cdots A^{\beta_k}$ Thus the coefficient of an explanatory variable appearing in a regression in logarithmic form, the corresponding coefficient is independent of the units $\ln A_i = x_i = x'_i + 1000$ $A_i = 1000 A'_i$

SIAI Swiss Institute of Artificial Intelligence	**Name: Hye-Young, Park** **Class: STA502: Math & Stat for MBA** **Date: 2021/09/06** **Problem Set 1 - Question 5**
QUESTIONS	***NOTES***
Question 5. **1.** Suppose that the true relationship between X and Y is linear, i.e. Y = β0 + β1X + ϵ. Consider the training residual sum of squares (RSS) for the linear regression, and also the training RSS for the cubic regression. Would we expect one to be lower than the other, would we expect them to be the same, or is there not enough information to tell? Justify your answer.	$RSS = \sum_{i=1}^{100} e_i^2 = \sum_{i=1}^{100} (y_i - \hat{y})^2$ A cubic regression model (Fitting Model): $Y = \beta_0 + \beta_1 X^1 + \beta_2 X^2 + \beta_3 X^3 + \varepsilon$ A linear regression (True Model): $Y = \beta_0 + \beta_1 X + \varepsilon$ 1. RSS for a cubic regression model $RSS = \sum_{i=1}^{100} e_i^2 = \sum_{i=1}^{100} (y_i - \beta_0 - \beta_1 X^1 - \beta_2 X^2 - \beta_3 X^3)^2$ 2. RSS for a linear regression model $RSS = \sum_{i=1}^{100} e_i^2 = \sum_{i=1}^{100} (y_i - \beta_0 - \beta_1 X)^2$ if $\beta_2 X^2 + \beta_3 X^3 > 0$ or $X > -\frac{\beta_2}{\beta_3}$, As seen above, in most case RSS for the cubic regression model is less than RSS for the linear regression model. However, the true model is the linear model, then β2 and β3 are close to 0. Thus the difference between two RSS would be trivial.
2. Suppose that the true relationship between X and Y is not linear, but we don't know how far it is from linear. Consider the training residual sum of squares (RSS) for the linear regression, and also the training RSS for the cubic regression. Would we expect one to be lower than the other, would we expect them to be the same, or is there not enough information to tell? Justify your answer.	In order to know which is higher or same, at least the information of sign(positive/negative) of β2 and β3 is needed.