Question 4. Consider the multiple linear regression model $y = X\beta + \epsilon$ with k explanatory variables in X.

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 $1/\lambda$. Use this result to explain what will happen when a particular explanatory variable is measured in thousands of kgs instead of millions of kgs.

   [Answer]

   The multiple linear regression model : $y = X\beta + \epsilon$
   Unfolding the model : $y = \beta_0 + \beta_1 x_1 + \beta_2 x_2 + \cdots + \beta_k x_k + \epsilon$

   Consider the particular explanatory variable are multiplied by λ
   : $y = \beta_0 + \beta_1 x_1 + \beta_2 x_2 + \cdots + (\frac{\beta_k}{\lambda})(\lambda x_k) + \epsilon$

   The particular explanatory variable is measured in thousands of kgs instead of millions of kgs.
   It means the unit of measure is reduced by 1/1000. So $\lambda = 1000$ and $\hat{\beta_k}$ also reduced by 1/1000.

   Then t-statistic for the significance of coefficient doesn't change :

   $$t - stat = \frac{\frac{1}{\lambda}\hat{\beta_k}}{\frac{1}{\lambda}s.e(x_k)} = \frac{\hat{\beta_k}}{s.e(x_k)}$$

   And the other explanatory variables, y, ϵ and R^2 also be same.

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.

   [Answer]

   Consider the particular explanatory variable are added by λ
   : $y = \beta_0 + \beta_1 x_1 + \cdots + \beta_k(x_k + \lambda) + \epsilon = (\beta_0 + \lambda\beta_k) + \beta_1 x_1 + \cdots + \beta_k x_k + \epsilon$

   In logarithmic form
   : $y = \beta_0 + \beta_1 \ln x_1 + \cdots + \beta_k \ln \lambda x_k + \epsilon = (\beta_0 + \lambda\beta_k) + \beta_1 \ln x_1 + \cdots + \beta_k \ln x_k + \epsilon$
   the corresponding coefficient is independent of the units.

   Then other explanatory variables, y, ϵ, R^2 and t-statistic doesn't change.
   The constant term is only changed : $(\beta_0 + \lambda\beta_k)$

Question 5. A researcher has collected a set of data (= 100 observations) containing a single predictor and a quantitative response. She then fits a linear regression model to the data, as well as a separate cubic regression, i.e., $Y = \beta_0 + \beta_1 X^1 + \beta_2 X^2 + \beta_3 X^3 + \epsilon$.

1. Suppose that the true relationship between X and Y is linear, i.e. $Y = \beta_0 + \beta_1 X + \epsilon$. 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.

   [Answer]

   If the true relationship between X and Y is linear, β_2 and β_3 will be zero.
   Then the RSS for the cubic regression will be $(y - \beta_0 - \beta_1 X)^2$ same as the RSS for the linear regression. So we would expect them to be the same.

   If $\hat{\beta}_2 = \hat{\beta}_3 = 0$, estimated RSS $= (y - \hat{\beta}_0 - \hat{\beta}_1 X)^2$ is equal to linear model, else RSS always decrease otherwise. Hence, the training RSS for the cubic regression always decrease or equal.

   Adding more variables into a linear model always reduces the sum of squares and in turn increases the R^2 value even if the added variables doesn't seem to be related to Y. Including insignificant terms will bring about over-fitting. So we consider the adjusted R-squared which defined :

$$R^2_{adj} = 1 - \frac{RSS/(n-1-p)}{TSS/(n-1)}$$

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.

   [Answer]

   If true relationship is non-linear, β_2 and β_3 will be significant.($\beta_2 \neq 0$, $\beta_3 \neq 0$)
   So we would expect the RSS for the linear regression to be lower than the RSS for the cubic regression.
   Because adding more terms into a linear model always reduces the sum of squares and in turn increases the r-squared value.