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

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.

   X_m: variables measured in millions of kgs
   X_t: variables measured in thousands of kgs
   β_m: regression coefficients measured in millions of kgs
   β_t: regression coefficients measured in thousands of kgs

   $$\begin{aligned}\widehat{y} &= X_m\beta_m + \epsilon, \quad \lambda = 1,000\\ &= \lambda X_m \frac{1}{\lambda}\beta_m + \epsilon\\ &= X_t\,\beta_t + \epsilon\end{aligned}$$

   The unit of $X_m\beta_m$ and $X_t\beta_t$ are the same as that of y and the error term.

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.

   $$\begin{aligned}y &= \alpha + (X+\lambda)\beta + \epsilon\\ &= \alpha + \lambda\beta + X\beta + \epsilon\\ &= (\alpha + \lambda\beta) + X\beta + \epsilon\\ &= \alpha_{new} + X\beta + \epsilon \quad \because \lambda\beta\ :\ const\end{aligned}$$

   Therefore, coefficient β doesn't change, but constant term(intercept term) α does.

	X	
Y	X	logX
Y	$linear$	$linear - log$
	$\widehat{Y} = \alpha + X\beta$	$\widehat{Y} = \alpha + (logX)\beta$
logY	$log - linear$	$log - log$
	$log\widehat{Y} = \alpha + X\beta$	$log\widehat{Y} = \alpha + (logX)\beta$

   Table 1: Four variables of logarithmic transformations

   In this case, regression model defined as linear-log model.

   $$y = \alpha + (logX)\beta + \epsilon$$

   u: scale of unit of measure

   $$\begin{aligned}y &= \alpha + log(uX)\beta + \epsilon\\ &= \alpha + (logu + logX)\beta + \epsilon\\ &= (\alpha + (logu)\beta) + (logX)\beta + \epsilon\\ &= \alpha_{new} + (logX)\beta + \epsilon \quad \because (logu)\beta\ :\ const\end{aligned}$$

   Therefore, the coefficients are independent of the units of measure.

Problem 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 (SSR) for the linear regression, and also the training SSR 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.

   Even if the true relationship between the population and a predictor variable is linear, it is not known whether the number of data is sufficient to estimate the parameters because the population itself has a distribution(but, in general, one dataset with 100 observations would not be enough to train the model). If the observations were sampled to bettrer fit a cubic regression than linear one, then the SSR may be less in the cubic regression model, unlike the true(linear) relationship.

   Even if the parameters for the distribution are unknown, if there is information about the distribution shape of the population and there are a sufficient number of observations, it can be said that the SSR of the linear model is more likely to be smaller than the SSR of the cubic model.

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 (SSR) for the linear regression, and also the training SSR 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.

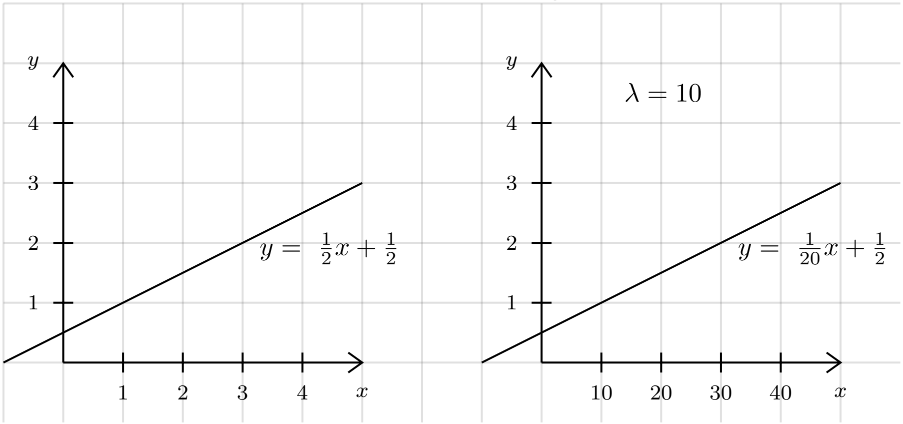
   As explained above, populations also have a distribution(but who knows). Therefore, if the number of observations is not sufficient, it is not certain which estimation model will have the lowest SSR. Of course, since it is not known how many observations are sufficient, it can only be expressed probability.

References

[1] Kenneth Benoit, Linear Regression Models with Logarithmic Transformations, 2011

[2] L. Tarkkonen, K. Vehkalahti, Measurement errors in multivariate measurement scales, Journal of Multivariate Analysis 96 (2005) 172–189

Problem 4.1 Example graph



Problem 5. Example graph