*Paper – what are the predictor(s) for the agriculture emission in a region*

*Shuyi Qi*

In order to predict the greenhouse gas emission accounted by the agriculture sector of a region, I combined data on the production quantities of five types of agricultural products in five different regions across the world between 2004 and 2014: *Crops Primary*, *Crops Processed*, *Live Animals, Livestock Primary* and *Livestock Processed* to fit a multiple regression model with all five production quantities as explanatory variables and the agriculture emissions of greenhouse gases (*CO2eq*, measured in gigagrams) in that region as the response. Among the explanatory variables, *Live Animals* is measured in million heads while the other four are measured in million tonnes. The five regions included in this dataset are: Africa, Americas, Asia, Europe and Oceania. The regression model has an of 99.93%, which means that 99.93% of the variability in the agriculture emission is explained by the five variables. The provides sufficient evidence that some part of the model is effective in predicting agriculture emissions of greenhouse gases. Below is the computer output for fitting :

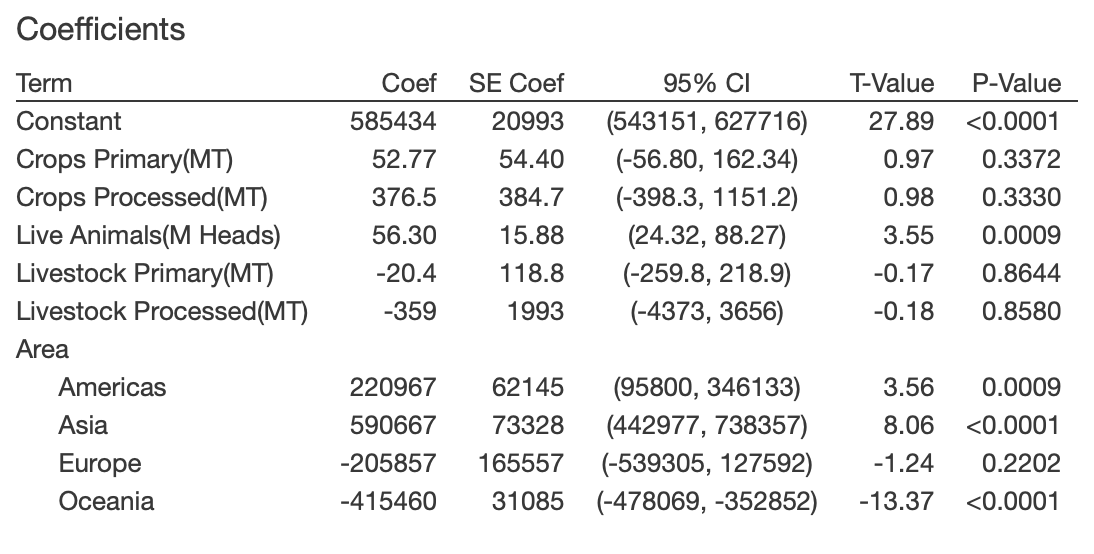


Table Multiple regression output for using Crops Primary, Crops Processed, Live Animals, Livestock Primary and Livestock Processed to predict CO2eq

We can see from *Table 1* that the coefficients of *Livestock Primary* and *Livestock Processed* have p-values of 0.8644 and 0.8580 respectively. These are very large p-values, which indicates that these two variables are not significant predictors in this model. This is because *Livestock Primary* and *Live Animals* are highly associated (r = 0.96, p-value 0), so if *Live Animals* is known, also knowing *Livestock Primary* does not contribute much additional information. This makes practical sense because in order to produce primary livestock products such as meat, milk, eggs, wool, silk and honey, more live animals need to be produced. Thus, variability in *Livestock Primary* is already reflected in the variability in *Live Animals*.

For *Livestock Processed*, if we try to fit a regression model with *Crops Primary, Crops Processed* and *Live Animals* as explanatory variables and *Livestock Processed* as the response, we obtain a model with a of 99.93%, which means that 99.93% of the variability in livestock(processed) is explained by the three variables all together. So, if *Crops Primary*, *Crops Processed* and *Live Animals* are known, knowing *Livestock Processed* does not provide much additional information. Hence, we can safely eliminate *Livestock Primary* and *Livestock Processed* without compromising the effectiveness of the model.

It turns out that by eliminating these two superfluous variables, we have an equally effective model, with a of 99.93%. Below is the Minitab output for fitting a model for *CO2eq* based on the three predictors:

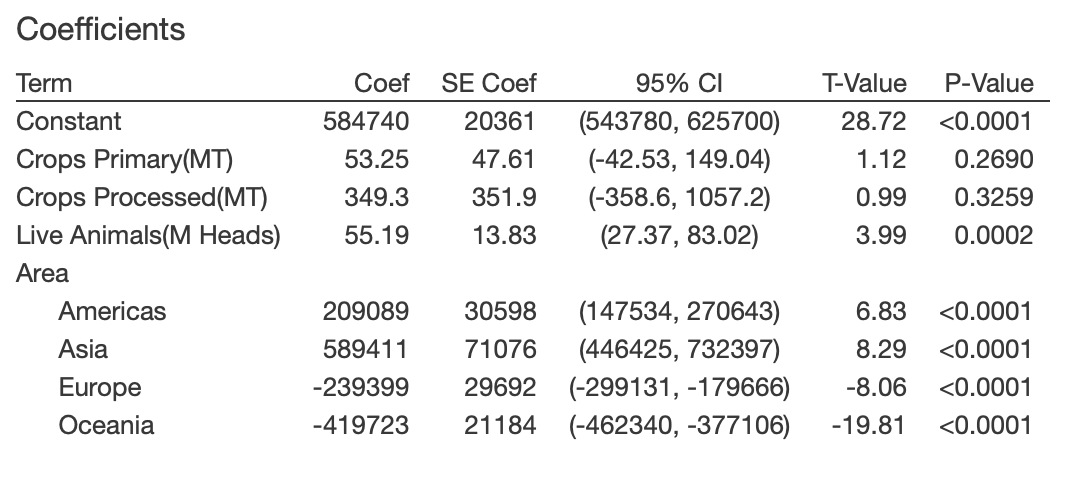


Table Multiple regression output using Crops Primary, Crops Processed and Live Animals to predict CO2eq

The adjusted increased by 0.01% after eliminating the two unnecessary variables. However, as we can see from *Table 2*, *Crops Processed* still has a relatively large p-value (0.3259), which indicates that it is not a significant predictor in this model. This is because *Crops Primary* and *Crops Processed* are highly associated (r = 0.96, p-value 0), so if *Crops Primary* is already in the model, also knowing *Crops Processed* becomes unnecessary and redundant. This makes practical sense because in reality, in order to produce more processed crop products like oils and wine, we will need to produce more primary crops like grains, fruits, vegetables and nuts and seeds. Thus, changes in the production of processed crop products will be reflected in the changes in the production of primary crop products. We can safely eliminate *Crops Processed* without compromising the effectiveness of the model. After eliminating *Crops Processed*, we now have a model with *Crops Primary* and *Live Animals* as explanatory variable and *CO2eq* as the response. The only decreases by 0.01% (from 99.93% to 99.92%) so we judge that removing *Crops Processed* is a good idea. The computer output for the new multiple regression model is shown below:

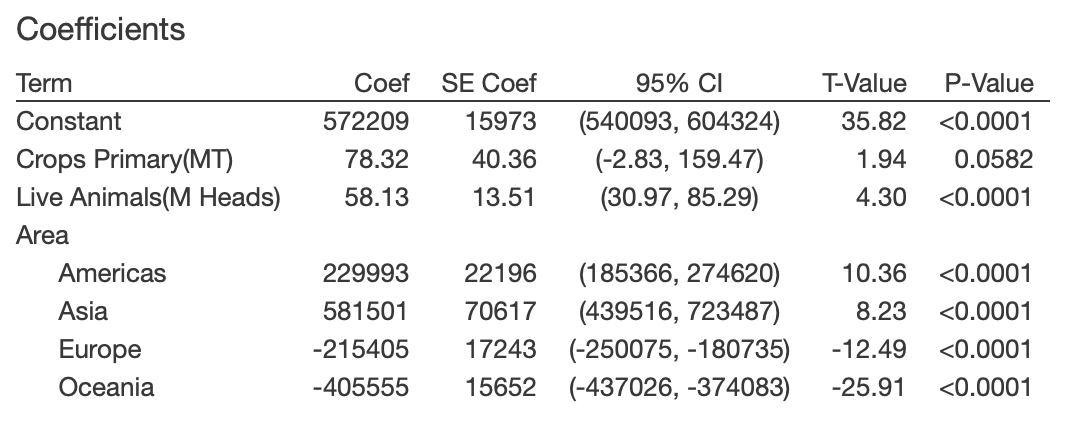


Table Multiple regression output using Crops Primary and Live Animals to predict CO2eq

In *Table 3*, we can see that the coefficient for *Crops Primary* has a p-value of 0.0582. This is a small p-value at a 10% significance level but still larger than 5%. We will compare the ANOVA tables for two models to determine whether *Crops Primary* is an effective predictor. Model A uses just a single predictor *Live Animals* while Model B adds *Crops Primary* as a second predictor to the model.

Model A:

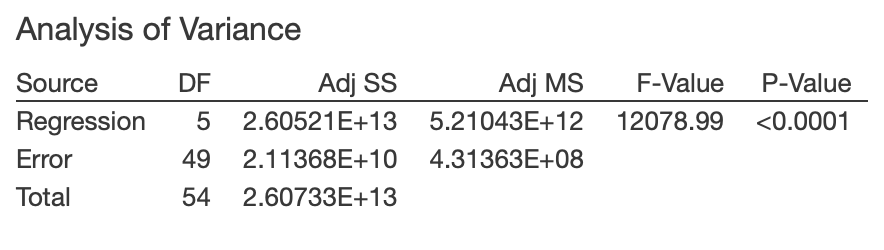


Table ANOVA table for fitting a model using only Live Animals to predict CO2eq

Model B:

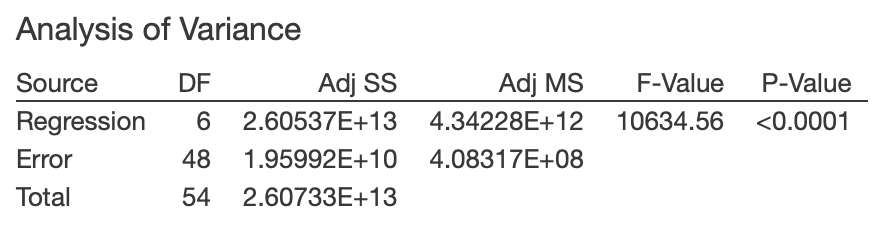


Table ANOVA table for fitting a model using both Crops Primary and Live Animals to predict CO2eq

The p-values in both *Table 4* and *Table 5* are close to zero, which gives strong evidence that at least one term in both models are effective for helping to explain *CO2eq*. We thus look at the of each model. It turns out that in both Model A and Model B, we have an of 99.92%. Therefore, eliminating *Crops Primary* does not reduce the effectiveness of the model, which means *Crops Primary* is not a helpful predictor as long as we have *Live Animals*. The reason could be that *Crops Primary* and *Live Animals* are also highly associated (r = 0.98, p-value 0). In reality, this positive association makes sense because a proportion of the primary crops are harvested as feed for animals. If the production of live animals increases, the production of primary crops has to increase to feed the extra animals. Thus, *Crops Primary* is more of an intermediate variable and becomes unnecessary after we have *Live Animals* as predictor. We would like the model to be effective and free of unnecessary or redundant variables, so we prefer Model A over Model B.

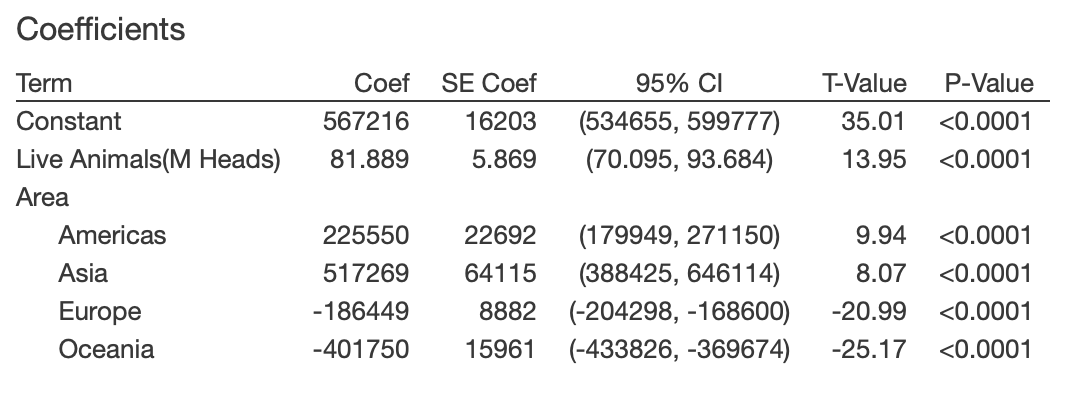


Table Multiple regression output using Live Animals to predict CO2eq

Now that we have decided on Model A, which has *Live Animals* as the only predictor and *CO2eq* as the response, we then want to verify the effectiveness of *Live Animals* as predictor. To do so, we perform a hypothesis test on the coefficient for *Live Animals.* We use vs . From *Table 4* and *Table 6* we see that t = 81.889/5.869 = 13.95, which gives a two-tailed p-value0 for a t-distribution with 55 – 5 – 1 = 49 degrees of freedom. This is a very small p-value, which means we have strong evidence that the coefficient of *Live Animals* differs from zero and that it is a useful predictor for *CO2eq* in this model. The Minitab output below shows the regression equations for predicting *CO2eq* with *Live Animals* in the five regions:

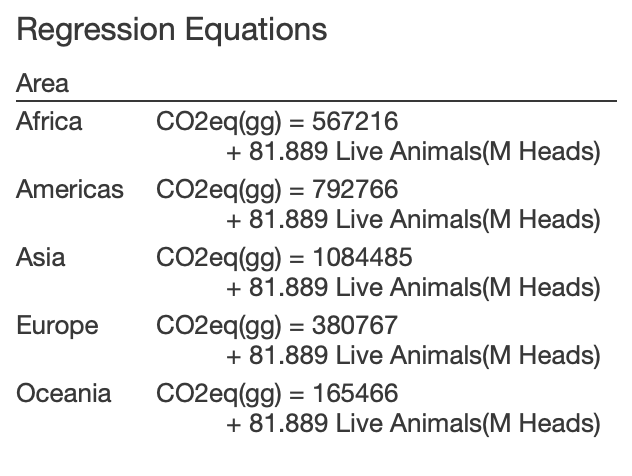


Figure Multiple regression equations for using Live Animals to predict CO2eq in each region

Before we can assert the effectiveness of the above regression model, we need to check the normality condition and the constant variability condition to verify that it is appropriate to use a regression model. To do so, we will look at its Residuals vs fits plot and histogram shown in *Figure 2* below:

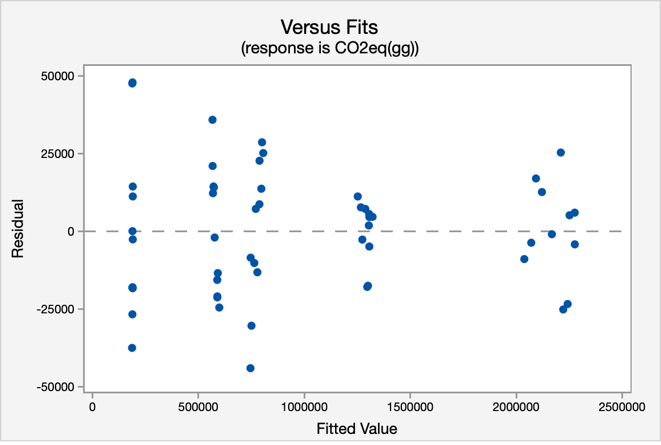
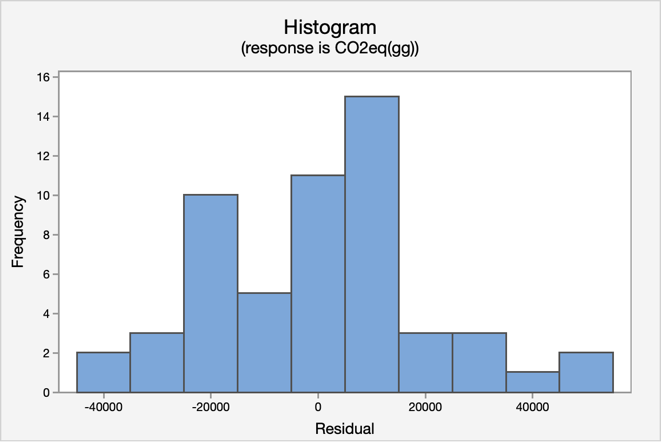


Figure 2 Residual plot with outliers

The residuals in the histogram in *Figure 2* are centered around zero. The distribution is relatively symmetric and bell-shaped, so the normality condition is reasonable. The residuals vs fitted values plot in *Figure 2* shows an even scatter on either side of the zero line. There is no sign of curvature. The computer output indicates that there are five significant outliers, three of which are positive while the other two are negative, which makes the variability appear first decreasing and then increasing. If we eliminate the outliers, the residuals vs fitted value plot shows a relatively constant variability as follows:

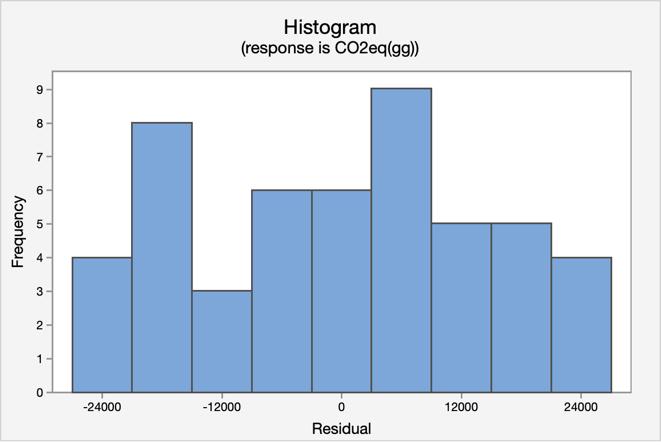
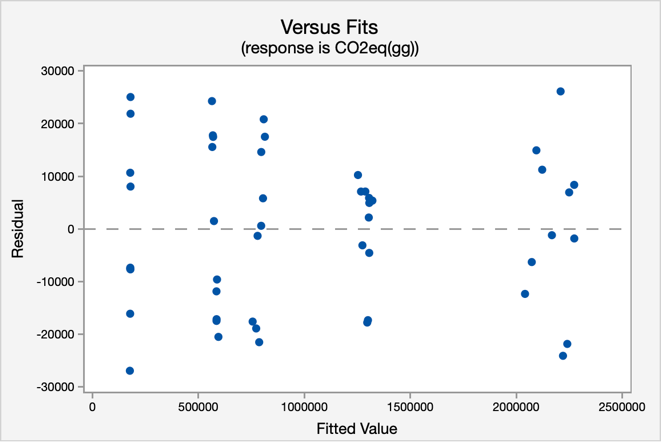
 

Figure 3 Residual plot without outliers

Based on *Figure 3*, we judge that there are no unusual patterns and that the constant variability is reasonably met. The five outliers correspond to Africa (2004), Europe (2004), Africa (2006), Oceania (2011) and Oceania (2012). Among these outliers, Africa (2004) and Africa (2006) have negative residuals and the other three have positive residuals. This implies that Africa in 2004 and 2006 has an observed greenhouse gas emission that is much lower than the expected while Europe in 2004 and Oceania in 2011, 2012 observe greenhouse gas emission much higher than expected. These outlier points are worth investigating so that we can figure out what factors contributed to their large residuals. We will eliminate the investigation for the purpose of this paper.

We have shown that the normality condition and the constant variability condition are met for the multiple regression model. Therefore, we can use this multiple regression model to predict the agriculture emission of a region. As we have seen above, we eliminated four of the five explanatory variables and the resulting regression model with *Live Animals* as the only explanatory variable is still effective. This implies that the process of producing *Live Animals* involves actions that emits a significant amount of greenhouse gases. Specifically, cows are the greatest contributors to the food-related greenhouse gas emissions. They release a significant amount of methane and CO2 every year. According to the Economist, if cows were a country, it would be the third largest greenhouse gas emitter. As the strong positive correlation between *Livestock Primary* and *Live Animals* reveals*,* a reduction in global demand for beef, milk and other dairy-based products will reduce the number of cows on our planet. The same relation applied to all other animal products like meat, eggs, wool, silk, and even honey. In our model, the coefficient of *Live Animals* is 81.889, which indicates that the *CO2eq* is expected to increase by 81.889 gigagrams with one million heads of increase in live animal production. On the other hand, *CO2eq* is also expected to decrease by 81.889 gigagrams with one million heads of decrease in live animal production. Therefore, a decrease in the demand for animal products will lead to a decrease in animal production as well as a decrease in global greenhouse gas emission.

In conclusion, by fitting a multiple regression model on the production quantities of each of the five agricultural products (primary crops, processed crops, live animals, primary livestock, processed livestock) to predict the amount of agricultural greenhouse gas emission, we find out that *Live Animals* alone explains 99.92% of the variability in the agricultural emissions of a region and that since the other four variables are highly associated with *Live Animals,* they turn out to be unnecessary and redundant in the model. This brings our attention to the effect of consuming animal products on global warming. A high demand in animal products leads to an increase in animal production, which leads to more crops harvested as feed to the animals, more grassland exploited for animal production and more methane and carbon dioxide emitted from the digestion system of cows. It follows that reducing the consumption of animal products is a good and effective way of slowing down global warming.