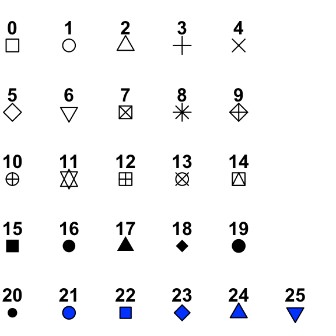
1. To save space, here is a graph for each numerical code and corresponding shape -



2. Looking at the intercept, it tells us that when the number of beaver damns is 0, the surface waters should be 606 hectare, with P-value of the T-test being extremely small, we should have little doubt on the placement of the intercept. As for the predictor variable “dams.n,” it has an estimated value of 0.318, meaning that the model predicts an increase of 1 unit of number of beaver dams would result in an increase of 0.318 hectare of surface water area. The P-value for this variable’s T-test is also significant (F =, p-value = 0.002718), so it is fairly a good fit to our data. The R2adj value is 0.57, meaning that our linear regression model explains 57% of the variability in the data, and I would claim that it is not doing a bad job, admittedly with more data points we might get a more precise predictions.

3. 

From the graphs shown above, it is quite interesting to observe the response variable and each individual predictor variable produce a different relationship. Overall (with broad generalization), the day of leaf out increases (later) as the latitude increases, and it decreases (sooner) as the max temperature increase. The general distribution for elevation vs. doy does not have obvious patters, yet there are no observations for plants to leaf-out after +120 days in the year at the elevations above 1000 meters. When it comes to the site environment, rural areas tend to have a higher/later leaf out days above 100, while in urban settings, days of leaf out are more prevalent below 100 days.

4. In this case, looking for multicollinearity means finding correlations between independent variables. On the plots provided, it is easy to spot 6 sub graphs that displayed either positive/negative correlations on the top right corner, meaning there would be underlying problems when counting latitude and temperatures as regressors at the same time. However, the correlations are understandable, as the latitude increases, naturally the max and min temperature of the area would decrease. On the other hand, if a region’s max temperature increases, the min temperature should follow the pattern as well. The variables we need to look out for in a regression model is – Lat + Tmax, Lat + Tmin, Tmax + Tmin.

5. In the line shown in question5, the *ifelse* statement takes 3 argument, the first is a logical condition, the second and third are returns for a true or false result if the logical evaluation. **Ifelse (Condition, A, B)** reads as: if **Condition** is True, then return **A**, else (meaning **Condition** is False) return **B**.

Since ifelse is as a binary return, in our case, whatever in the column siteDesc that is not “Urban” would result in an “0” in the urID column, and suburban would correspond to 0, just like rural.

6.

Top left - The points are be symmetrically distributed around the horizontal line in the latter plot, showing that linearity assumption is met.

Top right – The qq-plot points stay relatively close to the diagonal line, showing that the assumption for normality of the error distribution is met.

Bottom left – The line is relatively flat with even distributions of data point on the graph, meeting the assumption for homoscedasticity of errors.

7. Since we are using urID as a binary dummy variable, the interpretation to its coefficient should be –

After taking the effects of other independent variables into account (or – holding other independent variables roughly the same), the model predicts that leaf-out date in urban area would be around 6 days earlier than in rural areas.

8.

* When all the numerical predictors are 0 and in rural setting, the leaf-out day of year is 159 (early June), which is significant based on the extremely low P-value.
* While holding other independent variables, 1-degree Celsius increase in the max temperature results in roughly 3 days sooner of the leaf out, which is significant based on the extreme low P-value.
* While holding other independent variables, an in crease of 1000mm precipitation results in roughly 2 days later of the leaf out. However, based on a 0.68 P-value, there’s insufficient evidence in our sample to conclude that a non-zero correlation exists between days of leaf out and precipitation.
* While holding other independent variables, increasing the elevation by 100 meters would making the day of leaf out early by nearly 3 days, which is significant based on the extreme low P-value.
* While holding other independent variables, the leaf out date in urban area would be 6 days earlier than in rural areas, which is significant based on the extreme low P-value.
* There are 315 entries of data that do not have temperature and precipitation recorded (mainly in 2019) and they are omitted, so perhaps the model would have been different.
* The Adjusted R-squared is 0.6755, meaning our model can account for 67.55% of the variability of the data, which is not bad given the long timespan (between years).
* The F-test resulted in significance, demonstrating that the sample data we have did not happen by chance, and our variables fits the data well comparing to an intercept-only model (F = 483.9, p < 2.2e-16).

9. I believe our finding is coherent to the common knowledge of leaf out, as the lower elevation and wormer temperature helps the tree to go green earlier (just look at Clinton, NY vs. Miami, Florida lol). I did not have any prior knowledge of urban environment causing leafing to start earlier, but perhaps the differences in air composition (higher CO2 and such in urban area) or the fertilizations might play a role.

10.

https://github.com/guozhaosengzs/ENVDS/blob/master/activity4/activity\_4\_script.R