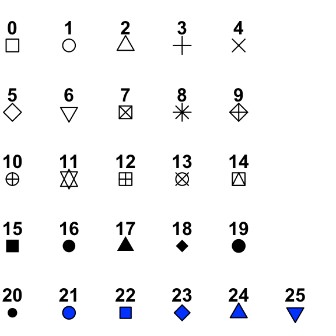
1. To save space, here is a graph for each numerical code and corresponding shape -



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 unit of surface water area. The P-value for this variable’s T-test is also significant (0.003), 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.

9.

10.