Multicollinearity

Multiple linear regression – Stat 230

In this class, you will explore the issue of multicollinearity, when two or more explanatory variables are highly correlated. The multiple linear regression model does not assume that the explanatory variables are independent, so multicollinearity is not a violation of our model assumptions, but it does cause problems with interpretations and inference for our coefficients.

Data overview

The owner of an apartment building in Minneapolis believed that their property tax bill was too high because of an over-assessment of the property's value by the city tax assessor. They hired an independent real estate appraiser to investigate the appropriateness of the city's assessment. In the original data set collected, there were five explanatory variables which the appraiser believed would be relevant to predicting sales price. However, when the appraiser reported their results to the owner of the apartment building, they noticed that the model did not take into account the level of the neighborhood amenities. As a good proxy for this rather hard-to-quantify feature, the appraiser decided to use the total payroll and and average salary in of residents in their analysis. Their argument was that the total payroll variable measures the level of commercial activity in the neighborhood, while the average salary variable is a measure of the affluence of neighborhood residents. Since these data are publicly available from the City of Minneapolis' Department of Community, Planning and Economic Development, the assessor records these data in the previous data set.

Variable names	Description
`price`	Sale price
`units`	Number of apartments
`age`	Age of structure (years)
`lot`	Lot Size (sq. ft)
`parking`	Number of On-Site Parking Spaces
`sqft`	Gross Building Area (sq. ft)
`payroll`	Total Payroll
`salary`	Average salary

You can load this data set using the command

```
mnsales <- read.csv("https://aloy.rbind.io/data/mnsales.csv")
```

Task 1. Fit a multiple regression model using all explanatory variables. Report the equation of the fitted model.

Task 2. Conduct an F-test of the overall utility of your multiple regression model (this is the `F-statistic` reported at the end of the `summary()` output). Make sure to state a null and an alternative hypothesis, the value of the test statistic, a p-value, and a conclusion within the context of the problem.

Task 3. In the daily prep, we discussed several indicators of multicollinearity. List them.

Task 4. Make the scatterplot matrix for all possible explanatory variables. Do any pairs seem highly correlated?

Task 5. Calculate the correlation matrix using the below code. Are any pairs highly correlated?

Task 6. Print the table of coefficients for your regression model. Do any of the estimated coefficient ***disagree*** with the pairwise correlations with weight? (That is, do any have different signs?)

Task 7. Looking again at the table of coefficients, which of the variables appear to be important predictors at the $\alpha = 0.05$ level? Do any of these results surprise you?

Task 8. Reflecting you your answers so far, are there any signs of multicollinearity?

Task 9. Another diagnostic tool used to detect multicollinearity is the variance inflation faction (VIF). To calculate the VIF values for each predictor, load the {car} package and run the following command (assuming that your model is called `mod1`). Are there any signs of multicollinearity based on the VIF values?

```
car::vif(full_mod)
```

Task 10. You should have found indications of multicollinearity in the previous task. Based on the stated goal of the analysis, should we be concerned about this? Why or why not?

Task 11 (If there's time) Regardless of your answer to the previous task, let's try to remedy the situation with multicollinearity. Propose a reduced model and check whether it resolves the issue with multicollinearity.