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 5 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")
```

What is this Adjusted R-squared that R keeps reporting?

Before we launch into our investigation of multicollinearity, let's discuss a common question that I have received about the summary output from a regression model in R: "what is this Adjusted R-squared value?"

Task 1. Fit three regression models using `price` as the response variable. For each, report both the the R^2 value and the adjusted R^2 values. Are they different? How do they change?

(a) Use only `sqft` as a predictor variable.
(b) Use only `sqft` and `age` as predictor variables.
(c) Use only `sqft`, `age`, and `parking` as predictor variables.

Recall that R^2 describes the proportion of variability in the response that is explained by the regression model. Adding a new predictor variable will only increase the amount of variability explained, or at the worst have no impact (if the added predictor is independent of the response). Consequently, R^2 is not a useful metric to compare fitted models.

The adjusted R^2, R^2_{adj}, imposes a *penalty for model complexity*, so it increases only if the improvement (i.e., increase in the explained variability) outweighs the cost of making the model more complex. The formula for R^2 is given by

$$R^2_{\text{adj}} = 1 - \left(\frac{n-1}{n-p-1}\right) \cdot (1 - R^2)$$

Notice that the fraction $(n-1)/(n-p-1)$ is the *penalty term* and is close to 1 if the model is small (i.e., if $p-1$ is close to 1).

So which version of R^2 should you report? If you are just looking at a single model or models all of the same size (i.e., same number of coefficients), then the "original" R^2 is fine. If you want to compare the fit of different size regression models, the use R^2_{adj}.

Detecting multicollinearity

Task 2. Fit a multiple regression model using all explanatory variables. Give the equation of the fitted model.

Task 3. Conduct an F-test of the overall utility of your multiple regression model. 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 4. In the daily prep, we discussed several indicators of multicollinearity. List them.

Task 5. For the full model, report the p-values associated with the individual $t-$tests of the seven parameters. At the $\alpha = 0.05$ level, which of the variables are significant? Do the estimated values for the seven parameters above match what the correlations between y and the respective

predictors suggest? Which ones do and which ones do not? To answer this question, fill in the following table:

You may find a scatterplot matrix of the data set to be useful:

```
GGally::ggpairs(mnsales, columns = c(2:8, 1))
```

Parameter	Estimate	p-value	Signif. (Y or N)	Corr.	Agreement
β_1					
β_2					
β_3					
β_4					
β_5					
β_6					
β_7					

Task 6. Based on the above table, are there any signs of multicollinearity?

Task 7. 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?

```
vif(mod1)
```

Task 8. 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 9 (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.