Multicollinearity: Sources and Assessment

Dr. Austin Brown

Kennesaw State University

Introduction

Some of today's materials were adapted from those created by Dr. Taasoobshirazi and my former professor, Dr. Khalil Shafie (thanks Drs. S & T!)

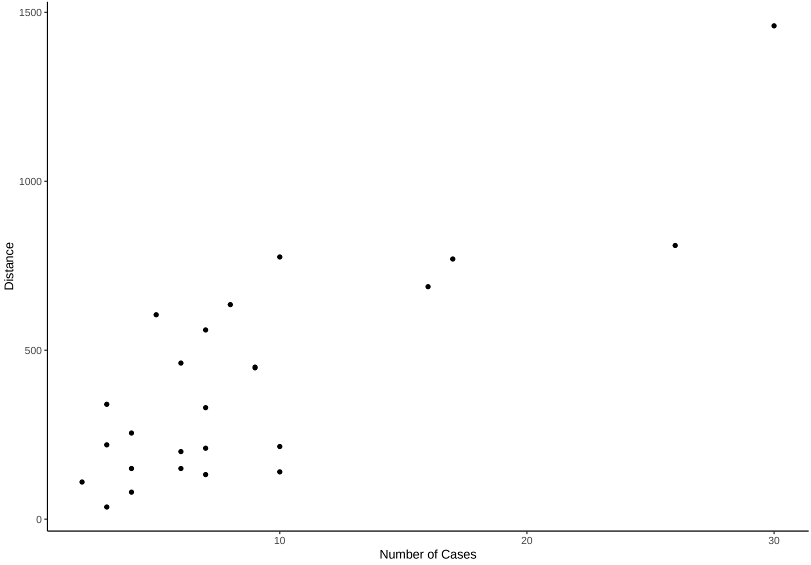
Introduction

- We have discussed in the prior couple of class sessions this additional assumption MLR models have: we want to have little to no *multicollinearity*.
- Multicollinearity is the interrelatedness the predictor variables have with one another.
- How do variables become interrelated with one another? Well, there are actually a few common causes:

Sources of Multicollinearity

- Multicollinearity can be the result of several things (polynomial models, models with interaction terms, and models with lots of categorical predictors will inherently have MC).
- First, the sampling technique being used could be a possible source (this is what your text calls it; I think a better term is "undercoverage").

Sources of Multicollinearity



Sources of Multicollinearity

- We've only sampled observations where number of cases and distance move in the same, positive direction.
- It's highly likely there are deliveries where the distance from the truck to the store/vending machine is small but the number of cases is great and vice versa.
  - We don't have that data.
- Because of this, our sample is almost certainly not representative of the population of delivery times as we're only examining a very specific subset.
  - We've introduced undercoverage sampling bias.

Sources of Multicollinearity

- Second, we can sometimes run into the same issue as undercoverage bias, except it isn't bias. It's just the nature of the relationship under consideration.

- For example, suppose we wish to build a model where residential energy consumption is being predicted by family income and home size.
  - Both would make sense to have as predictors!

- However, it is clear that income and home size almost certainly have some degree of positive correlation between them (i.e., one is sort of a proxy for the other), but this is a function of the research question being investigated, and not a result of undercoverage bias.

Sources of Multicollinearity

- Third is a source we've already discussed and one we will discuss: polynomial regression and interaction terms.
- Obviously, since adding higher order polynomial terms (new predictor variables which are functions of existing predictor variables, like x_1^2 and x_1^3) involves the creation of new variables which are functions of existing variables, a degree of dependency is inherent.
- The same thing happens when including interaction terms as we saw in the lecture on including categorical predictors.

Sources of Multicollinearity

- Finally, we can also run into issues of multicollinearity in instances where we have an overdefined or overfit model where there are a large number of predictors.
  - This is super common in medical research.

- In such cases, it may be valuable to either rely on existing research to help determine which subset of variables should be used or use a method like principal component analysis (PCA) where we can determine which subset of variables are important in model fit.

Issues Associated with Multicollinearity

- So far, we've talked about ways multicollinearity can occur in a regression model. But why do we care so much about it?
- Before we get into that, let's first discuss the most common measure of assessing the degree of multicollinearity called the *Variance Inflation Factor* or VIF for short.
- For each $\hat{\beta}_j$ in our regression model, we will have an associated VIF_j.

Issues Associated with Multicollinearity

- For a given predictor variable with associated $\hat{\beta}_j$:

$$VIF_j = \frac{1}{1 - R_j^2}$$

- where R_j^2 is the coefficient of determination for a model where the jth predictor serves as the outcome and the remaining $j-1$ predictors serve as predictors in this new model.
- We literally interpret VIF_j as the factor by which the variance for $\hat{\beta}_j$ increases due to multicollinearity.

Issues Associated with Multicollinearity

- In general, a VIF_j value exceeding 10 (which corresponds to $R_j^2 = 0.90$) is considered unacceptably high and corrective action ought to be taken.

- Okay this is well and good, but getting back to the original question, why does this matter? What problems does it cause which warrant all of this discussion?

- For starters, as we discussed in our conversation on categorical predictors, if we have a perfect linear combination of our predictors (where we can manipulate some of our predictors to exactly yield one of our others), then the $(X^T X)^{-1}$ matrix does not exist.
  - This means that we do not have unique estimates nor estimates with minimum variance for our vector of β estimates. See the Gauss-Markov theorem for why this is the case.

Issues Associated with Multicollinearity

- Second, note that the variance for a given $\hat{\beta}_j$ estimate is given by:

$$Var[\hat{\beta}_j] = \sigma^2 C_{jj} = \sigma^2 \frac{1}{1 - R_j^2} = \sigma^2 VIF_j$$

- And also recall that when we're performing a t-test for a single regressor, the t-test is:

$$t_0 = \frac{\hat{\beta}_j}{\sqrt{\hat{\sigma}^2 C_{jj}}} = \frac{\hat{\beta}_j}{\sqrt{\hat{\sigma}^2 VIF_j}}$$

Issues Associated with Multicollinearity

- So what does this mean? As $VIF_j \to$ big, $\implies t_0 \to 0$.
- As a result, this means that even if the alternative, $H_1 : \beta_j \neq 0$ is true, there's a VIF big enough for us to fail to reject H_0.
- Consequently, the probability of making a Type II error goes up and conversely, our statistical power goes down.
  - We'd obviously like to avoid this to the greatest degree possible!

Diagnostics for Multicollinearity

- We've already discussed using VIF as a way to detect multicollinearity, but there are some others we can also employ.

- One simply method of assessing multicollinearity is through the examination of the off-diagonal elements in the $X^T X$ matrix, denoted r_{ij}.

- These off-diagonal elements represent the pairwise correlation between x_i and x_j where absolute values of r_{ij} approaching 1 indicate a potential problem.

- This method isn't very effective, however, since it only considers pairwise dependency, and really isn't that different from a scatterplot matrix.

Diagnostics for Multicollinearity

- One interesting approach to assessing multicollinearity is by calculating two measures called the condition number and the condition indices of our $X^T X$ matrix.

- These measures are functions of the eigenvalues of the $X^T X$ matrix. Eigenvalues (denoted by λ) are special scalars which are a solution to the below linear system of equations (specific to square matrices).

$$\mathbf{A} \;\; = \lambda$$

- Without getting into the nuts and bolts too much, it can be shown that the product of a square matrix's eigenvalues is equal to its determinant (which remember, is sort of a measure of a matrix's variability and must be non-zero in order for a square matrix to be invertible).

Diagnostics for Multicollinearity

- What this suggests is, if we have strong linear dependency, between our predictors, one (or more) of our eigenvalues has to be around zero.

- Thus, the condition number is the maximum eigenvalue divided by the minimum eigenvalue.

$$\kappa = \sqrt{\frac{\lambda_{max}}{\lambda_{min}}}$$

- If $\kappa < 10$, we don't have a problem. $10 \leq \kappa \leq 30$ indicates a mild to moderate problem. $\kappa > 30$ indicates a severe problem.

Diagnostics for Multicollinearity

- The condition indices are:

$$\kappa_j = \sqrt{\frac{\lambda_{max}}{\lambda_j}}, \quad j = 1, 2, \dots, p$$

- If several κ_j's exceed about 30, then this indicates that we have lots of issues with multicollinearity.

- Let's see how we can calculate these using R.

Solutions to Multicollinearity

- Especially in working with the Acetylene data, we could see a big problem with multicollinearity with all of the methods we learned about.
- So now the question is: how do we deal with it?
    - The easiest way is to throw variables out. But this isn't always prudent!!
- In three weeks, we will learn about two modern methods, LASSO and Ridge, which we can use to correct this problem.