Collinearity

2024-10-15

Symptoms of collinearity

1) Collinearity between independent variables

- High r^2 values between X-variables
- Statistically-significant relationships between X-variables

2) High variance inflation factors (VIF) of variables in model
3) Variables significant in simple regression, but not in multi-variable regression
4) Individual variables not significant in multi-variable regression model, but the overall multi-variable regression model is significant
5) Large changes in coefficient estimates between full and reduced models
6) Large SE in multi-variable regresion models, despite high power

Simulation exercise 1

- I simulated 1,000 datasets with varying degrees of collinearity (correlation) between two X-variables. Here is truth:
  - Simulations: $n = 1,000$
  - $y = 10 + 3X_1 + 3X_2 + \epsilon \sim N(0, 2)$ – both X variables have effects on Y!
  - $X_1 = U[0, 10]$
  - $X_2 = X_1 + N(0, z)$
  - For each simulation, I used a different value of z from a uniform distribution: $z = U[0.5, 20]$.

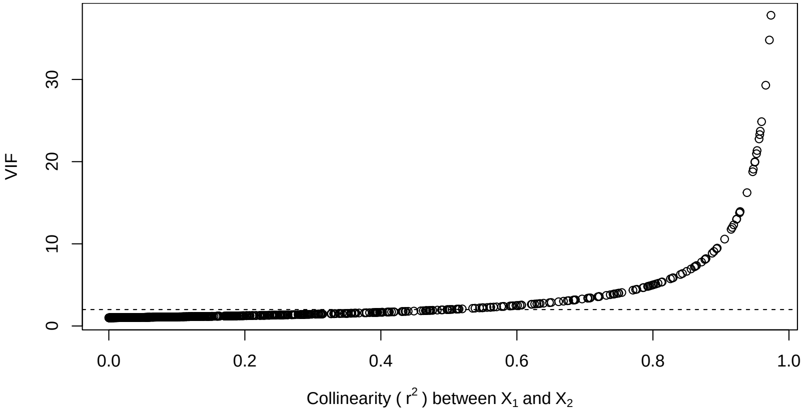
Methods

For each simulation, I did a few things:

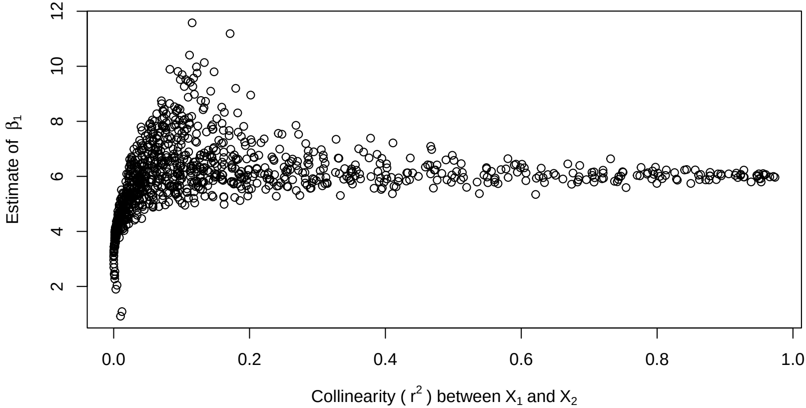
- Fit a **simple model** ($Y \sim X_1$) and measured the estimate, SE, and p-value for $\beta 1$
- Fit a **multi-variable model** ($Y \sim X_1 + X_2$) that included both of the collinear, confounding variables, and measured the effect, SE, and p-value for $\beta 1$.
- Measured how collinearity between X_1 and X_2 (i.e., r^2) influenced the the **Variance Inflation Factor** from the multi-variable model

Variance Inflation Factor

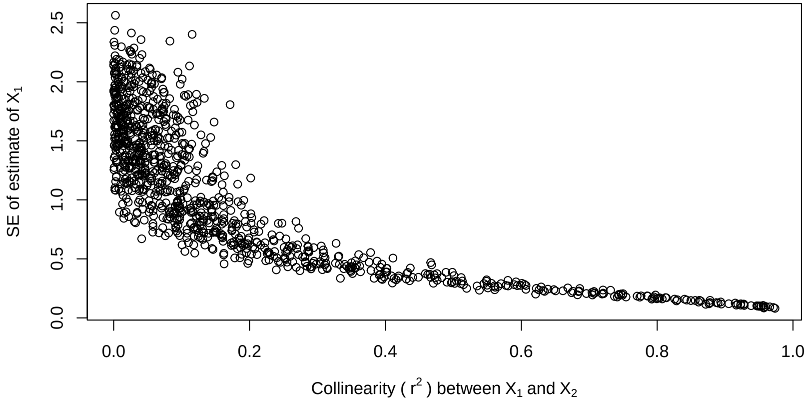
Variance Inflation Factor (VIF) – the amount (in *times*) that the variance (SE^2) in the β increases due to collinearity



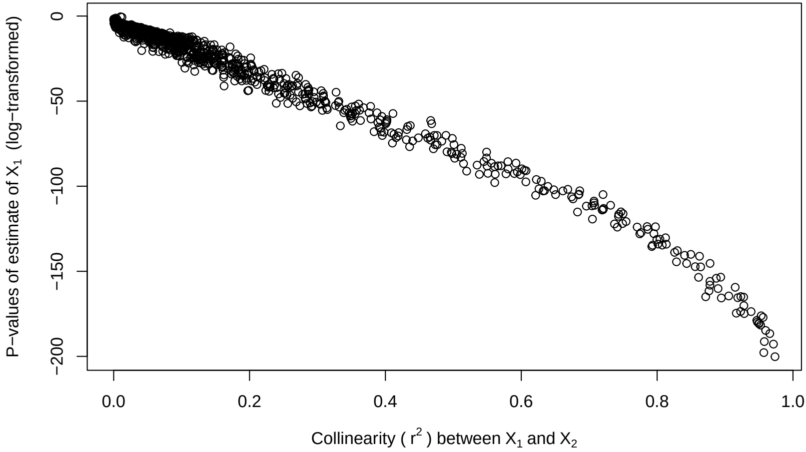
Simple model: $Y \sim X_1$



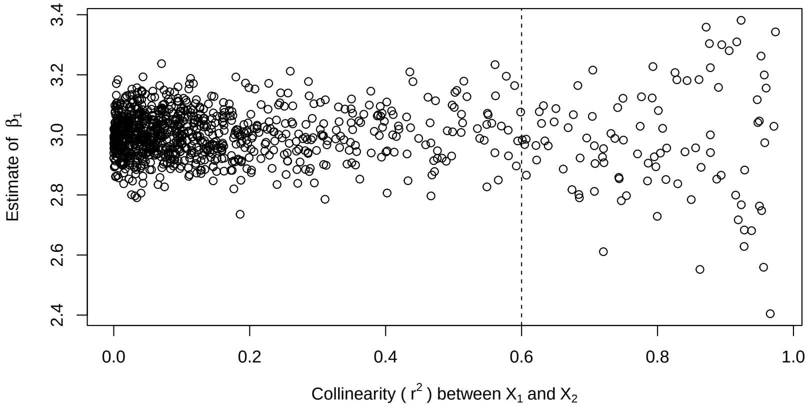
Simple model: $Y \sim X_1$



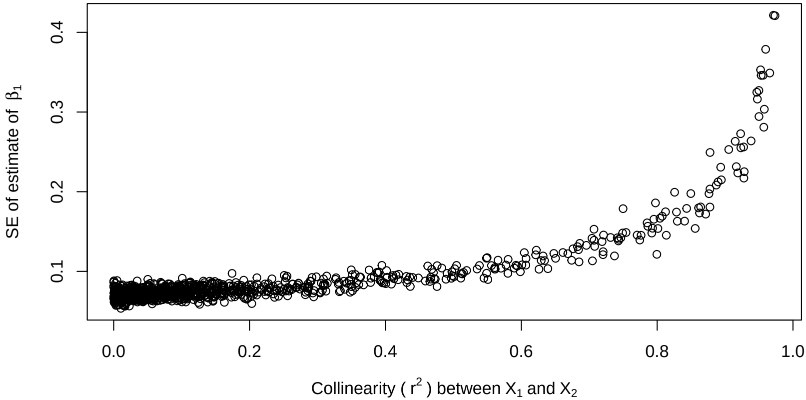
Simple model: $Y \sim X_1$



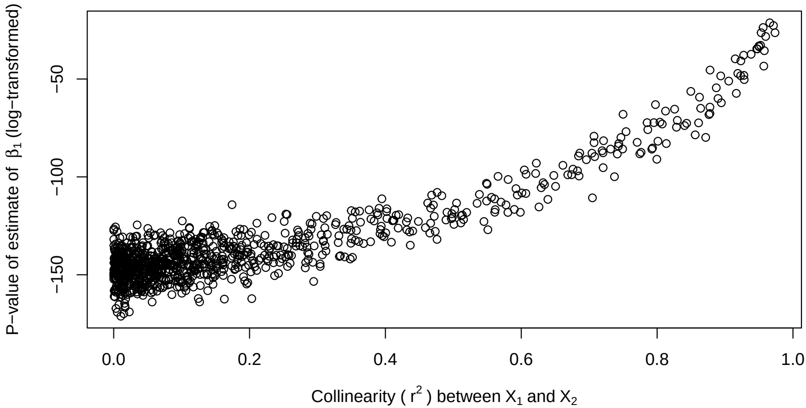
Multi-variable model: $Y \sim X_1 + X_2$



Multi-variable model: $Y \sim X_1 + X_2$



Multi-variable model: $Y \sim X_1 + X_2$



Confounding variables

Confounding variable – a variable that will bias results if you leave it out.

- Correlated with another X-variable
- Has it's own effect on Y

To avoid negative effects of confounding variables, I recommend:

1) **Sample in a manner that eliminates collinearity**.
2) **Use multi-variable regression**.
3) **Include confounding variables, even if they are non-significant**.
4) **Get more data!** This decreases SE and VIF.

Redundant variables

Redundant variables – collinear X-variables that don't have an effect on the Y-variable.

- Correlated to another X-variable, but
- Do not have an effect on Y-variable

A useful way to think about confounding or redundant variables might be with the βs.

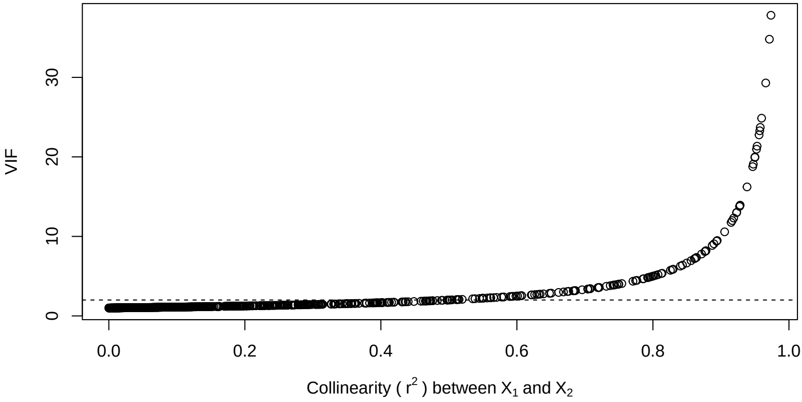
- If the $\beta \neq 0$, it's a confounding variable.
- If the $\beta = 0$, it's a redundant variable.

Simulation exercise 2

This simulation exercise is similar as before, but now only X_1 has an effect, and X_2 is a redundant variable.

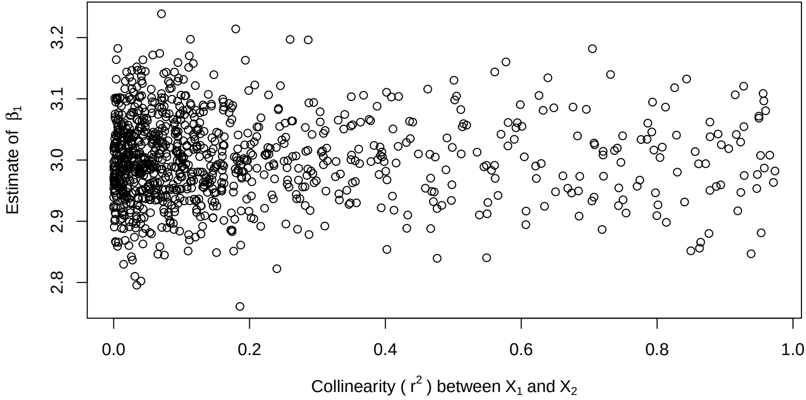
- $X_1 = 3$
- $X_2 = 0$

Variance Inflation Factor

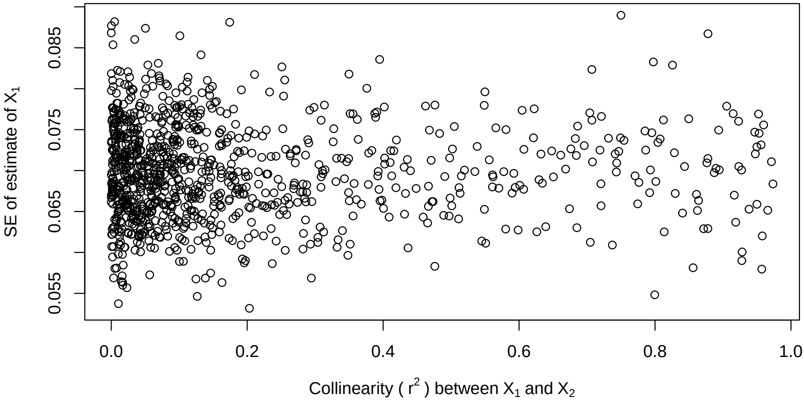


Simple regression model: $Y \sim X_1$

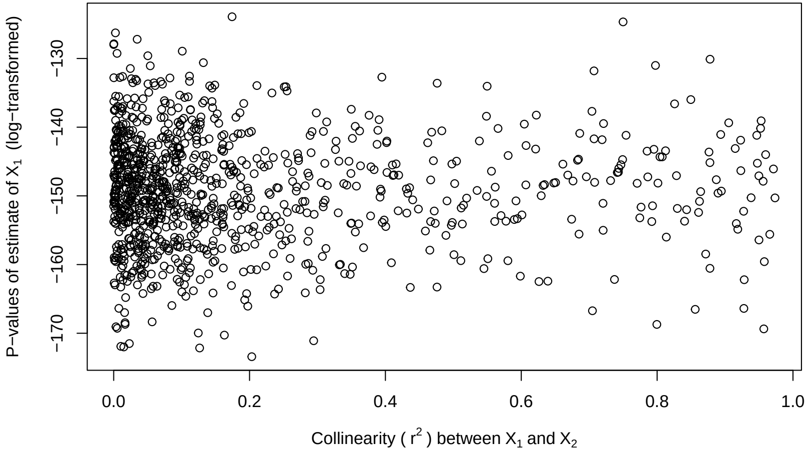
Now, we see that the existence of a redundant variable does not influence our estimation of β_1 using a simple linear model! This is good.



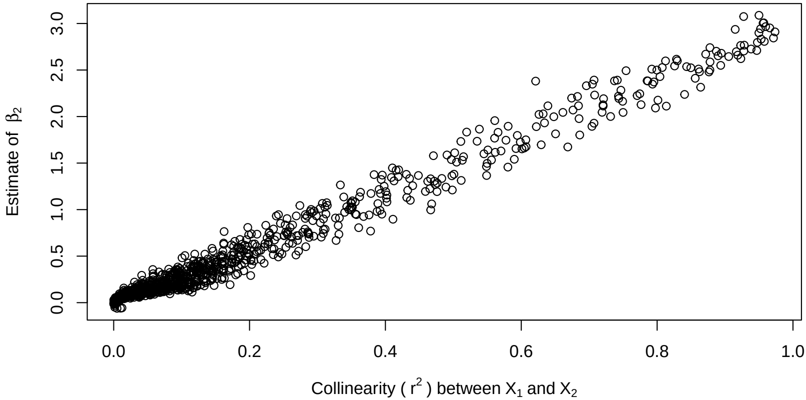
Simple regression model: $Y \sim X_1$



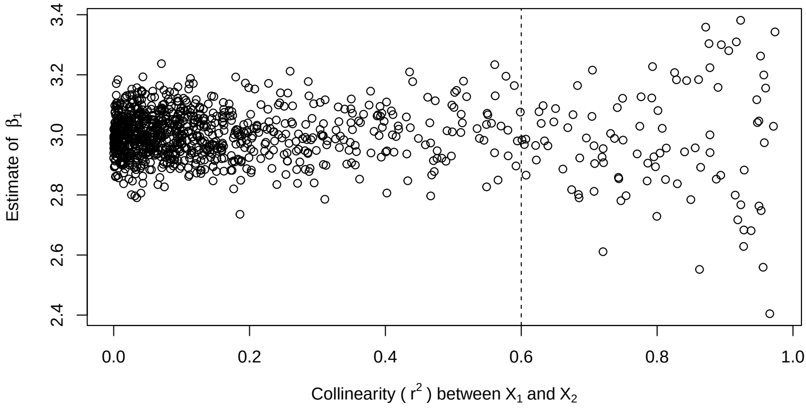
Simple regression model: $Y \sim X_1$



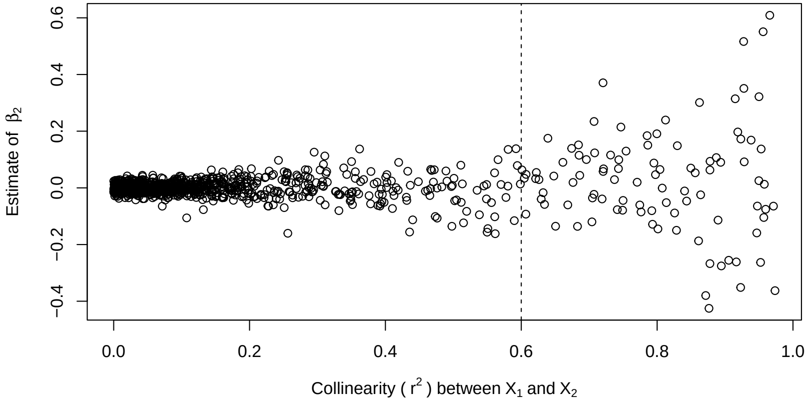
Simple regression model w/ redundant variable: $Y \sim X_2$



Multi-variable model: $Y \sim X_1 + X_2$



Multi-variable model: $Y \sim X_1 + X_2$



Practical guidance for examining and dealing with collinearity

Do you have collinearity in your data or system?

1) Be careful to identify potential confounding variables prior to data collection. Use logic and try to identify all confounding variables and measure these.
2) Calculate collinearity and VIF among independent variables – before you start your analysis. High collinearity between X-variables tends to imply redundancy.
3) Pay attention to how coefficient estimates and variable significance change as variables are removed or added.

Practical guidance for examining and dealing with collinearity

Is a variable redundant or confounding?

1) Think! Use logic.
2) If there is extreme collinearity, there's likely a **redundant** variable
3) Large changes in coefficient estimates of **both variables** between full and reduced models: variables are likely **confounding**.
4) Large changes in coefficient estimates of **one variable** between the reduced and full model, and the full model estimates a variable to be close to zero: **redundant**
5) Not sure whether it's redundant or confounding? **Assume confounding & include it.** Multi-variable regression also produces unbiased estimates (on average) regardless of the type of collinearity.

Practical guidance for examining and dealing with collinearity

What to do with **redundant variables**?

1) Determine which variable best explains the response using P-values from regression and changes in coefficient estimates with variable addition and removal
2) Do not include redundant variable in final model (to reduce VIF)

What to do with **confounding variables**?

1) Sample in a manner that eliminates collinearity, which can be due to real collinearity or sampling artifact.
2) Use multi-variable regression; may have large SE if collinearity is strong.
3) Include confounding variables, even if non-significant.
4) Get more data! Decrease SE due to variance inflation.