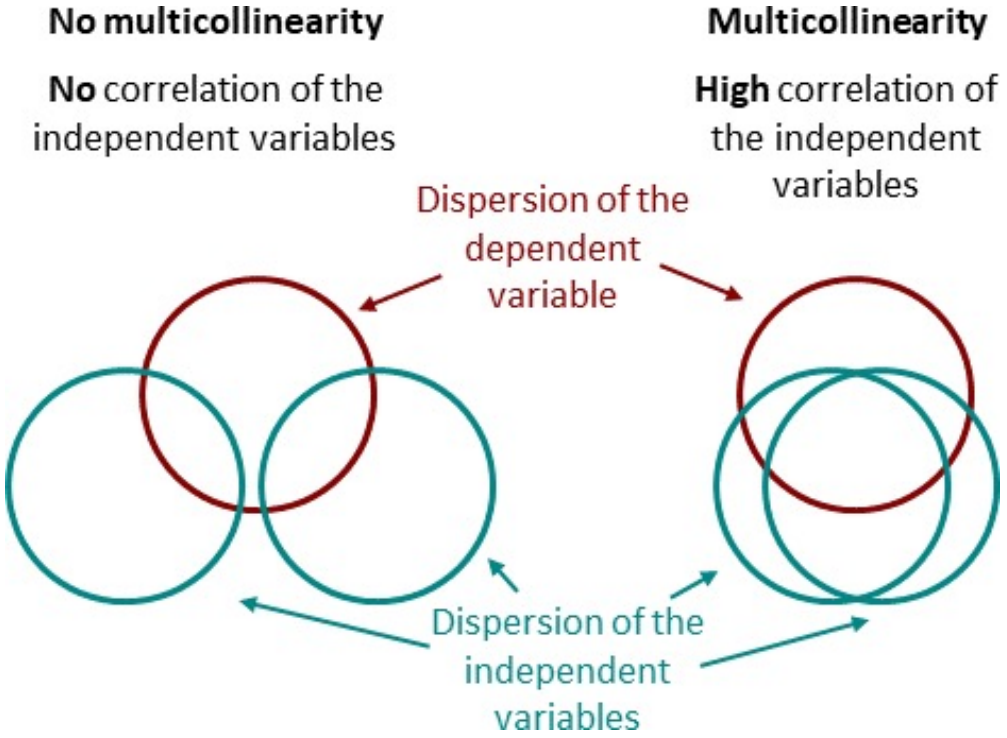
MULTICOLLINEARITY AND THE CHALLENGES OF BEING TOO SIMILAR

By Madeleine Armer

AGENDA

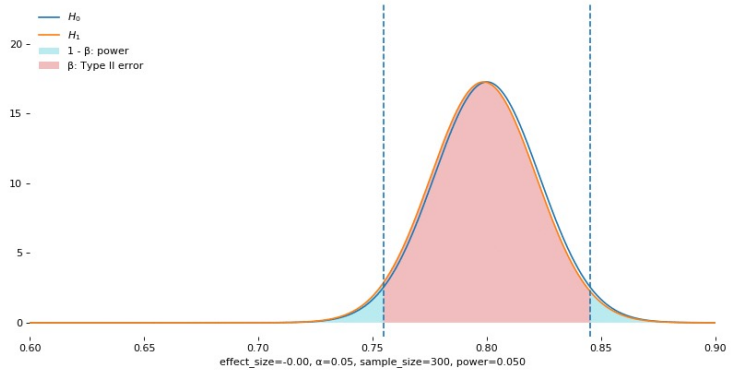
- Background on multicollinearity
- Issues of multicollinearity
- How to diagnose and fix
- Example problem

WHAT IS MULTICOLLINEARITY



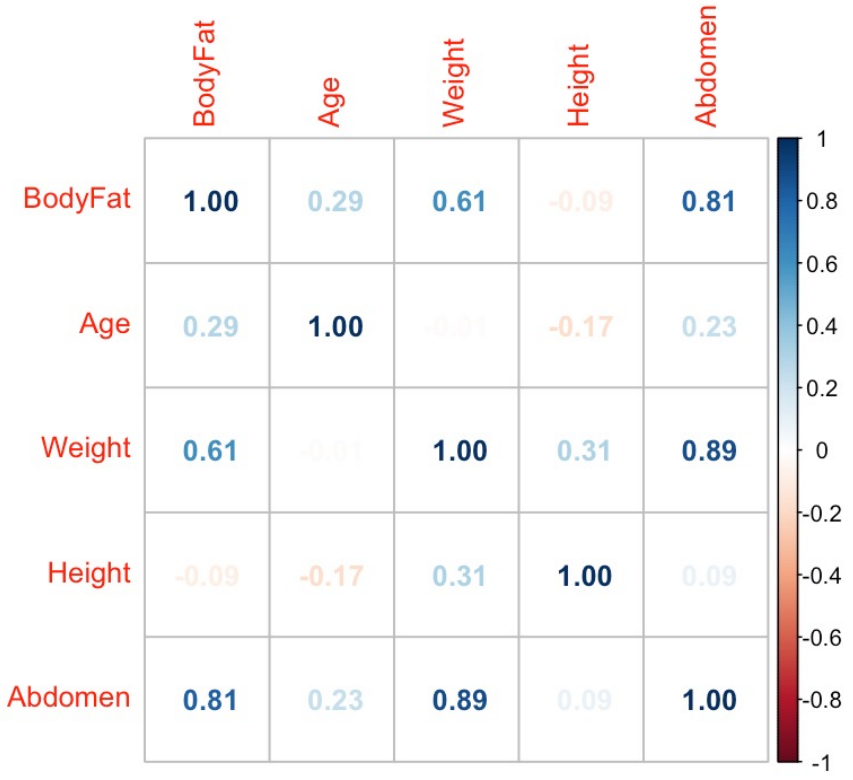
- Where predictors in a regression model are correlated to each other
- Assumptions of linear regression - independent predictors
- Too many variables doing the same thing – lack of uniqueness

ISSUES WITH MULTICOLLINEARITY



- Increases standard errors of the affected variables
- Increases probability of type II errors in significance of predictors
  - H_0: $\beta_i = 0$
- Coefficients become unreliable for interpretation
  - β_1 explains the effects of X_1 on Y, and if X_1 and X_2 are correlated, now β_1 is also explaining X_2 effects as well

DIAGNOSTIC TOOLS



- EDA: Correlation plot
  - r > 0.80
- High overall Rsq value and many variables with no significance
  - low t-values and high pvalues

DIAGNOSTIC TOOLS

$$\text{VIF} = \frac{1}{(1 - R_i^2)}$$

- R_i^2 is the r-squared value of that predictor regressed on all other predictors in the model
- $1 - R_i^2$ is Tolerance -- coefficient of non-association
- VIF = 1 – 4 
- VIF > 4 

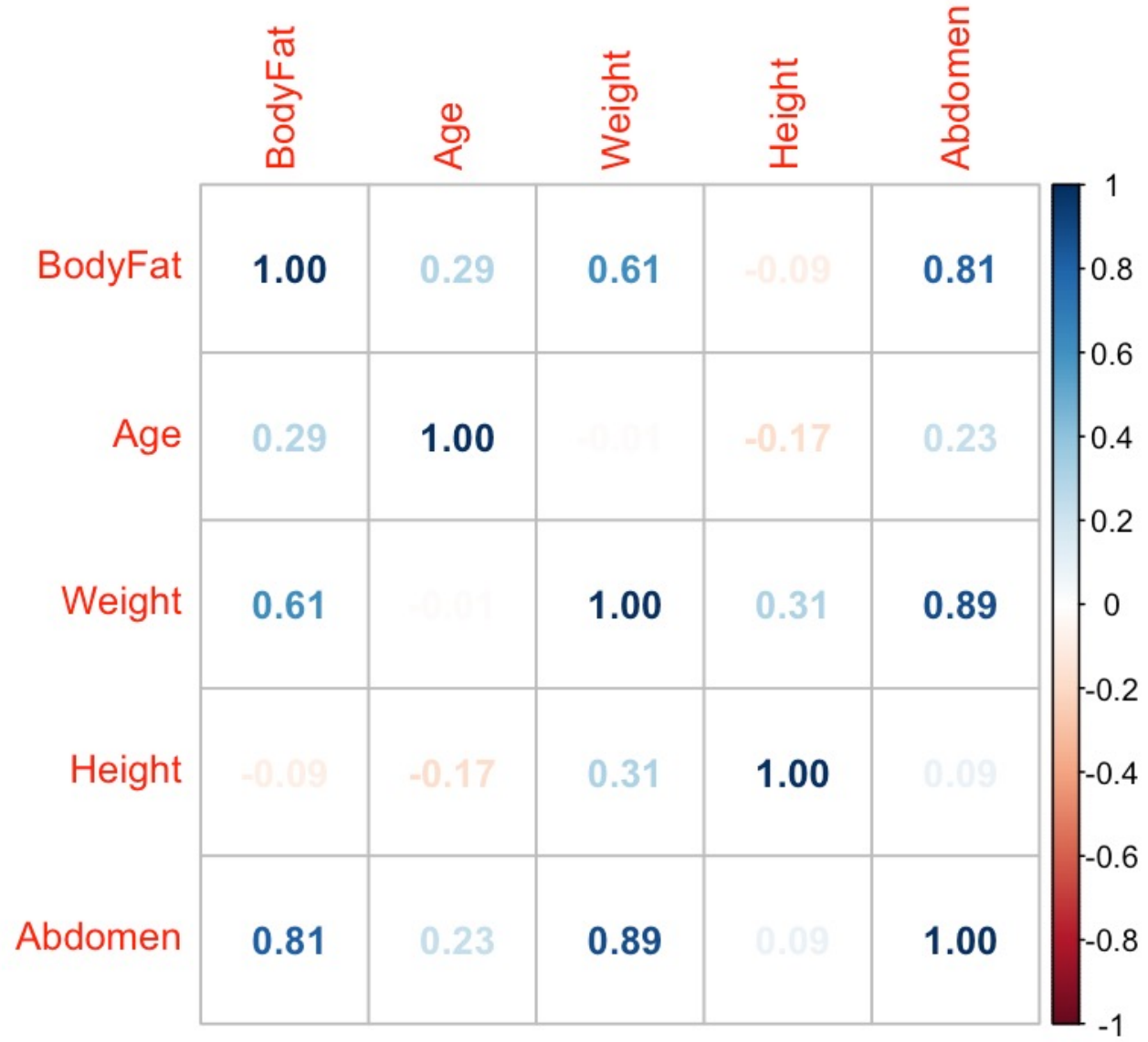
HOW TO FIX IT

- Remove one of the highly correlated variables
- PCA or PLS regression
- Stepwise selection functions
- Or ignore it! 😉

WHEN YOU CAN IGNORE IT

- Between levels of a categorical dummy variable
- Polynomial regression
  - $(X + X^2 + X^3)$
- Predictive models – It depends …
  - When your coefficients are not of interest

EXAMPLE:


BodyFat
Age
Weight
Height
Abdomen
BodyFat 1.00 0.29 0.61 -0.09 0.81
Age 0.29 1.00 -0.01 -0.17 0.23
Weight 0.61 -0.01 1.00 0.31 0.89
Height -0.09 -0.17 0.31 1.00 0.09
Abdomen 0.81 0.23 0.89 0.09 1.00
1
0.8
0.6
0.4
0.2
0
-0.2
-0.4
-0.6
-0.8
-1

EXAMPLE:

```
Call:
lm(formula = Age ~ BodyFat + Weight + Height + Abdomen, data = bodyfat1)

Residuals:
     Min       1Q   Median       3Q      Max
-28.9090  -7.5088  -0.5427   7.7818  23.9083

Coefficients:
            Estimate Std. Error t value Pr(>|t|)
(Intercept) -16.47062   17.97857  -0.916    0.360
BodyFat      -0.03477    0.15402  -0.226    0.822
Weight       -0.47351    0.06165  -7.681 3.70e-13 ***
Height        0.20360    0.21670   0.940    0.348
Abdomen       1.43115    0.21090   6.786 8.49e-11 ***
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 10.78 on 247 degrees of freedom
Multiple R-squared:  0.2793,    Adjusted R-squared:  0.2677
F-statistic: 23.94 on 4 and 247 DF,  p-value: < 2.2e-16
```

```
> vif(model)
  BodyFat    Weight    Height   Abdomen
 3.585389  7.084584  1.359749 11.161462
```

EXAMPLE CONT.

```
Call:
lm(formula = Age ~ BodyFat + Weight + Height, data = bodyfat1)

Residuals:
    Min      1Q  Median      3Q     Max
-30.725  -8.270  -0.803   8.762  32.381

Coefficients:
            Estimate Std. Error t value Pr(>|t|)
(Intercept) 63.79967   14.71746   4.335 2.12e-05 ***
BodyFat      0.69151    0.12039   5.744 2.70e-08 ***
Weight      -0.12023    0.03589  -3.350 0.000935 ***
Height      -0.15177    0.22858  -0.664 0.507338
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 11.72 on 248 degrees of freedom
Multiple R-squared:  0.145,     Adjusted R-squared:  0.1346
F-statistic: 14.02 on 3 and 248 DF,  p-value: 1.801e-08
```

```
> vif(model2)
 BodyFat   Weight   Height
1.854011 2.032303 1.280341
```

REFERENCES:

- https://www3.nd.edu/~rwilliam/stats1/x91.pdf
  - This one is great resource for the theory behind how SE is affected
- https://analystprep.com/study-notes/cfa-level-2/multicollinearity/
- https://datatab.net/tutorial/multicollinearity
- https://blog.minitab.com/en/understanding-statistics/handling-multicollinearity-in-regression-analysis
- https://corporatefinanceinstitute.com/resources/knowledge/other/variance-inflation-factor-vif/
- https://www.analyticsvidhya.com/blog/2020/03/what-is-multicollinearity/

THANK YOU FOR LISTENING!

Questions??

Preparing people to lead extraordinary lives