Huddle Presentation – Multicollinearity

8/23/2022

MULTICOLLINEARITY

1. What is it?

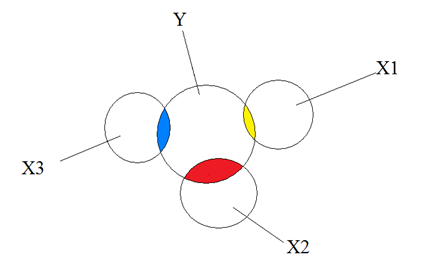
When two or more covariates in a model are highly correlated.

When one covariate is linearly expressed by other covariates in the model (a linear function)

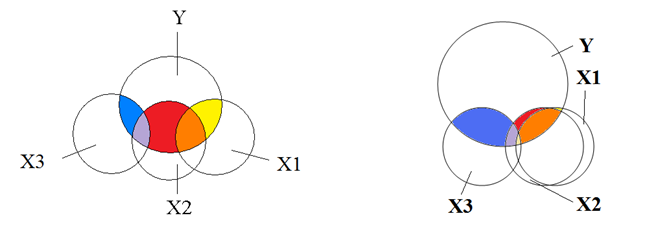
Y = X1 + X2 + X3

Red, blue and yellow are where each X explains variance in Y

No multicollinearity:



Moderate: Extreme:



[What is Multicollinearity? A Visual Description (theanalysisfactor.com)](https://www.theanalysisfactor.com/multicollinearity-explained-visually/)

1. Is it a problem? May not affect the accuracy of the model but might make interpreting the effects for individual predictors problematic.
2. Can lead to skewed or misleading results when interpreting independent variable(s)
   1. Large standard errors -- wide confidence intervals -- less reliable probabilities
   2. Imprecise or nonsensical estimates: 0 or huge
3. Suspect when:
   1. Large standard errors
   2. Regression coefficients change when predictor variables are added or deleted
   3. Regression coefficients not significant in a multivariable regression model but significant in a simple linear regression (unexpected results)
   4. Coefficients opposite of what was expected (unexpected results)
   5. R2  for the model is high but no/few significant predictors
4. Detect using:
   1. Variance inflation factor (VIF) - includes all variables in the model
   2. Correlation –bivariate relationship between two variables
   3. Condition indices2
5. Reduced by:
   1. Remove variable(s) or combine or transform (mean centering, standardize)
   2. Ridge regression or partial squares regression or principal component regression

VIF

1. Represents how well the variable is explained by the other independent variables.1
2. Measures the strength of the correlation between predictor variables. Regression model is fit using each independent variable (IV) as the dependent and all other variables as predictors

X1 = X2, X3, X4 -> R2

R2 indicates the percentage of variance in the individual IV that the set of IVs explains

X2 = X1, X3, X4 -> R2

X3 = X1, X2, X4 -> R2

X4 = X1, X2, X3 -> R2

General rule of thumb

1 = not correlated

2-4 = moderately correlated. May/not require attention.

5+ = highly correlated\*

\* some use 2.5 others 10

A VIF of 4.0 means that the standard error for the coefficient of that predictor variable is 4.0 times larger than if that predictor variable had 0 correlation with the other predictor variables.

(Source: Statology)

STATA EXAMPLE: LINEAR REGRESSION

global labs VARLIST

global labr REVISED\_VARLIST

asdoc regress adl\_difficulty\_num $labs, title(Table 1: Regression results) save(H:\Stats\multicollinearity\results tables.rtf) replace

asdoc vif, label title(Table 2: Multicollinearity ) save(H:\Stats\multicollinearity\results tables.rtf) append

asdoc pwcorr VARIABLES, label save(H:\Stats\multicollinearity\results tables.rtf) title(Table 3: Pearson Correlations) append

asdoc regress adl\_difficulty\_num $labr, title(Table 4: Revised regression results with variables removed) save(H:\Stats\multicollinearity\results tables.rtf) append

asdoc vif, label title(Table 5: Multicollinearity revised) save(H:\Stats\multicollinearity\results tables.rtf) append

Table 1: Regression results

|  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- |
| adl\_difficulty\_num | Coef. | St.Err. | t-value | p-value | [95% Conf | Interval] |
| msas\_phys\_num\_pres~b | -1.245 | 1.659 | -0.75 | .507 | -6.524 | 4.034 |
| mortality\_score | -.018 | .131 | -0.14 | .899 | -.436 | .400 |
| eligibility\_bun\_10 | -.023 | .064 | -0.36 | .741 | -.226 | .180 |
| eligibility\_sbp\_10 | -.050 | .129 | -0.39 | .725 | -.459 | .360 |
| eligibility\_cr\_10 | .108 | .332 | 0.33 | .766 | -.947 | 1.1630 |
| mean\_life\_expectancy | .238 | .379 | 0.63 | .574 | -.966 | 1.443 |
| heart\_failure\_adm~10 | .655 | .700 | 0.94 | .418 | -1.573 | 2.883 |
| age\_10 | -.212 | 1.293 | -0.16 | .880 | -4.326 | 3.903 |
| bun\_units1\_11 | .089 | .106 | 0.84 | .462 | -.248 | .426 |
| pain\_bother\_b | .925 | 1.246 | 0.74 | .512 | -3.041 | 4.892 |
| nausea\_bother\_b | .270 | .983 | 0.27 | .802 | -2.86 | 3.400 |
| kccq\_domain\_100 | .097 | .106 | 0.92 | .425 | -.239 | .433 |

Table 2: Multicollinearity

|  |  |  |
| --- | --- | --- |
|  | VIF | 1/VIF |
| bun units1 11 | 9.228 | .108 |
| MSAS num physical symptoms | 9.225 | .108 |
| kccq domain 100 | 9.072 | .11 |
| SBP at time of study eligibility | 7.267 | .138 |
| Seattle heart failure score | 5.702 | .175 |
| Number of heart failure admits | 5.453 | .183 |
| Mean life expectancy | 4.472 | .224 |
| Pain bother b | 3.047 | .328 |
| age 10 | 2.989 | .335 |
| Creatinine | 2.259 | .443 |
| BUN | 2.135 | .468 |
| Nausea bother b | 2.012 | .497 |
| Mean VIF | 5.239 | . |

Table 3: Pearson Correlations

|  |  |  |  |  |  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- |
| Variables | (1) | (2) | (3) | (4) | (5) | (6) | (7) | (8) | (9) | (10) | (11) |
| (1) msas\_phys\_num\_~b | 1.000 |  |  |  |  |  |  |  |  |  |  |
| (2) mortality\_score | -0.064 | 1.000 |  |  |  |  |  |  |  |  |  |
| (3) eligibility\_b~10 | 0.011 | 0.157\* | 1.000 |  |  |  |  |  |  |  |  |
| (4) eligibility\_s~10 | 0.064 | -0.030 | 0.046 | 1.000 |  |  |  |  |  |  |  |
| (5) eligibility\_c~10 | 0.016 | 0.008 | 0.277\* | 0.045 | 1.000 |  |  |  |  |  |  |
| (6) mean\_life\_expe~y | -0.019 | -0.405\* | -0.349\* | 0.206\* | -0.014 | 1.000 |  |  |  |  |  |
| (7) heart\_failure~10 | 0.168\* | 0.072 | -0.054 | -0.016 | -0.103\* | -0.064 | 1.000 |  |  |  |  |
| (8) age\_10 | -0.157\* | 0.137\* | 0.224\* | 0.059 | 0.034 | -0.257\* | -0.221\* | 1.000 |  |  |  |
| (9) bun\_units1\_11 | 0.121 | 0.189\* | 0.756\* | 0.201\* | 0.177\* | -0.366\* | -0.030 | 0.204\* | 1.000 |  |  |
| (10) pain\_bother\_b | 0.260\* | -0.071 | -0.062 | 0.050 | -0.065 | 0.050 | 0.234\* | -0.130 | -0.019 | 1.000 |  |
| (11) nausea\_bother\_b | 0.214 | 0.344\* | 0.180 | -0.048 | -0.009 | -0.072 | -0.045 | -0.120 | -0.051 | 0.008 | 1.000 |
| (12) kccq\_domain\_100 | 0.534\* | -0.002 | -0.070 | 0.038 | 0.034 | 0.029 | 0.152\* | -0.211\* | 0.002 | 0.202\* | 0.308\* |

Table 4 Revised: Regression results **-** bun\_units1\_11 removed from model

|  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- |
| adl\_difficulty\_num | Coef. | St.Err. | t-value | p-value | [95% Conf | Interval] |
| msas\_phys\_num\_pres~b | -.038 | .284 | -0.13 | .894 | -.621 | .545 |
| mortality\_score | -.025 | .023 | -1.11 | .275 | -.072 | .021 |
| eligibility\_bun\_10 | -.002 | .023 | -0.11 | .916 | -.051 | .046 |
| eligibility\_sbp\_10 | .006 | .025 | 0.25 | .803 | -.044 | .057 |
| eligibility\_cr\_10 | .139 | .145 | 0.96 | .347 | -.158 | .436 |
| mean\_life\_expectancy | -.025 | .116 | -0.22 | .831 | -.263 | .213 |
| heart\_failure\_adm~10 | .237 | .179 | 1.32 | .197 | -.103 | .604 |
| age\_10 | .021 | .362 | 0.06 | .954 | -.722 | .764 |
| pain\_bother\_b | .393 | .368 | 1.07 | .296 | -.363 | 1.148 |
| nausea\_bother\_b | .350 | .394 | 0.89 | .382 | -.459 | 1.159 |
| kccq\_domain\_100 | .010 | .018 | 0.54 | .594 | -.027 | .046 |

Table 1: Regression results – FOR COMPARISON

|  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- |
| adl\_difficulty\_num | Coef. | St.Err. | t-value | p-value | [95% Conf | Interval] |
| msas\_phys\_num\_pres~b | -1.245 | 1.659 | -0.75 | .507 | -6.524 | 4.034 |
| mortality\_score | -.018 | .131 | -0.14 | .899 | -.436 | .400 |
| eligibility\_bun\_10 | -.023 | .064 | -0.36 | .741 | -.226 | .180 |
| eligibility\_sbp\_10 | -.050 | .129 | -0.39 | .725 | -.459 | .360 |
| eligibility\_cr\_10 | .108 | .332 | 0.33 | .766 | -.947 | 1.1630 |
| mean\_life\_expectancy | .238 | .379 | 0.63 | .574 | -.966 | 1.443 |
| heart\_failure\_adm~10 | .655 | .700 | 0.94 | .418 | -1.573 | 2.883 |
| age\_10 | -.212 | 1.293 | -0.16 | .880 | -4.326 | 3.903 |
| bun\_units1\_11 | .089 | .106 | 0.84 | .462 | -.248 | .426 |
| pain\_bother\_b | .925 | 1.246 | 0.74 | .512 | -3.041 | 4.892 |
| nausea\_bother\_b | .270 | .983 | 0.27 | .802 | -2.86 | 3.400 |
| kccq\_domain\_100 | .097 | .106 | 0.92 | .425 | -.239 | .433 |

Table 5 Revised : Multicollinearity - variable removed from model

|  |  |  |
| --- | --- | --- |
|  | VIF | 1/VIF |
| Nausea bother b | 2.082 | .480 |
| MSAS num physical symptoms prese | 1.932 | .518 |
| age 10 | 1.739 | .575 |
| Seattle heart failure score (fro | 1.621 | .617 |
| Mean life expectancy ate eligibi | 1.577 | .634 |
| BUN at time of stidy eligibility | 1.576 | .634 |
| kccq domain 100 | 1.550 | .645 |
| Pain bother b | 1.458 | .686 |
| Creatinine at time of study elig | 1.340 | .746 |
| Number of heart failure admissio | 1.335 | .749 |
| SBP at time of study eligibility | 1.219 | .821 |
| Mean VIF | 1.584 | . |

STATA OPTIONS FOR DICHOTOMOUS & CATEGORICAL COVARIATES:

1. Chi-square test or Kappa agreement
2. Tetrachoric or polychoric correlations

asdoc tabulate lack\_of\_energy\_present\_b lack\_of\_appetite\_present\_b, chi nokey column save(H:\Stats\multicollinearity\results tables.rtf) title(Table 5: Energy & Appetite) append

asdoc, text(Pearson chi2 = 58.5677 Pr = 0.000) save(H:\Stats\multicollinearity\results tables.rtf) append

asdoc kap lack\_of\_energy\_present\_b lack\_of\_appetite\_present\_b, save(H:\Stats\multicollinearity\results tables.rtf) append

tetrachoric DICHOTOMOUS\_VARIABLES, star(0.05) pw

matrix tet = r(Rho)

putexcel set " H:\Stats\multicollinearity\results tables.rtf \correl table.xlsx", sheet(Tetrachoric) replace

putexcel A3 = matrix(tet), names nformat(number\_d2)

polychoric CATEGORICAL\_VARIABLES

matrix poly = r(R)

putexcel set " H:\Stats\multicollinearity\results tables.rtf \correl table.xlsx ", sheet(Polychoric) modify

putexcel A3 = matrix(poly), names nformat(number\_d2)

Table 6: Energy & Appetite

|  |  |  |  |
| --- | --- | --- | --- |
| Lack\_of\_energy\_present\_b | Lack\_of\_appetite\_present\_b | | |
| 0 | 1 | Total |
| 0 | 150 | 10 | 160 |
|  | 50.00 | 8.85 | 38.74 |
| 1 | 150 | 103 | 253 |
|  | 50.00 | 91.15 | 61.26 |
| Total | 300 | 113 | 413 |
|  | 100.00 | 100.00 | 100.00 |

First row has *frequencies* and second row has *column percentages*

Pearson chi2 = 58.5677 Pr = 0.000

|  |
| --- |
| Expected Agreement agreement Kappa Std. err. Z Prob>Z  61.26% 44.90% 0.2969 0.0388 7.65 0.0000 |

Table 7: Energy & Concentration

|  |  |  |  |
| --- | --- | --- | --- |
| Lack\_of\_energy\_present\_b | Diff\_concentrating\_present\_b | | |
| 0 | 1 | Total |
| 0 | 150 | 10 | 160 |
|  | 48.39 | 9.71 | 38.74 |
| 1 | 160 | 93 | 253 |
|  | 51.61 | 90.29 | 61.26 |
| Total | 310 | 103 | 413 |
|  | 100.00 | 100.00 | 100.00 |

First row has *frequencies* and second row has *column percentages*

Pearson chi2 = 48.7354 Pr = 0.000

|  |
| --- |
| Expected Agreement agreement Kappa Std. err. Z Prob>Z  58.84% 44.36% 0.2602 0.0373 6.98 0.0000 |

Table 8: Tetrachoric Correlations between Dichotomous Variables

|  |  |  |  |  |  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- |
|  | pain | energy | sob | appetite | nausea | drowsy | Dry mouth | constip | Diff sleep | Diff  concent | Weight loss |
| pain | 1.00 |  |  |  |  |  |  |  |  |  |  |
| Lack of energy | 0.40 | 1.00 |  |  |  |  |  |  |  |  |  |
| Sob | 0.30 | 0.57 | 1.00 |  |  |  |  |  |  |  |  |
| Lack of appetite | 0.41 | 0.66 | 0.46 | 1.00 |  |  |  |  |  |  |  |
| nausea | 0.32 | 0.45 | 0.38 | 0.49 | 1.00 |  |  |  |  |  |  |
| drowsy | 0.33 | 0.57 | 0.46 | 0.43 | 0.52 | 1.00 |  |  |  |  |  |
| Dry mouth | 0.26 | 0.34 | 0.31 | 0.35 | 0.25 | 0.31 | 1.00 |  |  |  |  |
| constipation | 0.10 | 0.14 | 0.05 | 0.26 | 0.34 | 0.28 | 0.28 | 1.00 |  |  |  |
| Diff sleeping | 0.32 | 0.44 | 0.31 | 0.47 | 0.52 | 0.39 | 0.35 | 0.18 | 1.00 |  |  |
| Diff concentrating | 0.33 | 0.63 | 0.34 | 0.50 | 0.49 | 0.46 | 0.44 | 0.25 | 0.48 | 1.00 |  |
| Weight loss | 0.19 | 0.20 | 0.15 | 0.45 | 0.33 | 0.23 | 0.13 | 0.17 | 0.08 | 0.21 | 1.00 |

STATA OTHER OPTIONS

Collin VARLIST – compute VIF, condition index, Eigen values

coldiag2 VARLIST – compute condition index

References:

1[What is Multicollinearity? Here’s Everything You Need to Know | by Aniruddha Bhandari | Analytics Vidhya | Medium](https://medium.com/analytics-vidhya/what-is-multicollinearity-heres-everything-you-need-to-know-b2e2e26108d4)

2[HG Notes-Identification of Multicollinearity-VIF and Conditioning Number\_20140304 (insead.edu)](https://www.insead.edu/sites/default/files/assets/faculty-personal-site/hubert-gatignon/documents/HG%20Notes-Identification%20of%20Multicollinearity-VIF%20and%20Conditioning%20Number_20140304.pdf)

[Eight Ways to Detect Multicollinearity - The Analysis Factor](https://www.theanalysisfactor.com/eight-ways-to-detect-multicollinearity/)

[What is Multicollinearity? A Visual Description (theanalysisfactor.com)](https://www.theanalysisfactor.com/multicollinearity-explained-visually/)