Advanced Statistics
DS2003 (BDS-4A)
Lecture 23

Instructor: Dr. Syed Mohammad Irteza

Assistant Professor, Department of Computer Science, FAST

12 May, 2022

Previous Lecture

- Adding More Explanatory Variables
- Interpreting a Linear Regression Model
- More on multicollinearity
  - Why is Multicollinearity a Potential Problem?
  - Types of Multicollinearity
  - Do I Have to Fix Multicollinearity?
  - Keep three points in mind

Testing for Multicollinearity with Variance Inflation Factors (VIF)

- Statistical software calculates a VIF for each independent variable. VIFs start at 1 and have no upper limit.
  - A *value of 1 indicates* that there is no correlation between this independent variable and any others.
  - *VIFs between 1 and 5 suggest* that there is a moderate correlation, but it is not severe enough to warrant corrective measures.
  - *VIFs greater than 5* represent critical levels of multicollinearity where the coefficients are poorly estimated, and the p-values are questionable.

What is Variance Inflation Factors (VIF)

- As the name suggests, a *variance inflation factor* (*VIF*) quantifies how much the variance is inflated.
- But what *variance*?
- Recall that we learned previously that the standard errors — and hence the variances — of the estimated coefficients are *inflated* when *multicollinearity* exists.
- A variance inflation factor exists for *each of the predictors* in a multiple regression model.

What is Variance Inflation Factors (VIF)

- For example, the *variance inflation factor* for the estimated regression coefficient b_j —denoted VIF_j —is just the factor by which the variance of b_j is "*inflated*" by the existence of correlation among the predictor variables in the model.

- In particular, the variance inflation factor for the j_{th} predictor is:

$$VIF_j = 1/(1 - R_j^2)$$

Where R_j^2 is the R^2-value obtained by regressing the j^{th} predictor on the remaining predictors.

How did we calculate R^2?

- $R^2 = 1 - \frac{Var(errors)}{Var(Y)}$

Pt	BP	Age	Weight	BSA	Dur	Pulse	Stress
1	105	47	85.4	1.75	5.1	63	33
2	115	49	94.2	2.1	3.8	70	14
3	116	49	95.3	1.98	8.2	72	10
4	117	50	94.7	2.01	5.8	73	99
5	112	51	89.4	1.89	7	72	95
6	121	48	99.5	2.25	9.3	71	10
7	121	49	99.8	2.25	2.5	69	42
8	110	47	90.9	1.9	6.2	66	8
9	110	49	89.2	1.83	7.1	69	62
10	114	48	92.7	2.07	5.6	64	35
11	114	47	94.4	2.07	5.3	74	90
12	115	49	94.1	1.98	5.6	71	21
13	114	50	91.6	2.05	10.2	68	47
14	106	45	87.1	1.92	5.6	67	80
15	125	52	101.3	2.19	10	76	98
16	114	46	94.5	1.98	7.4	69	95
17	106	46	87	1.87	3.6	62	18
18	113	46	94.5	1.9	4.3	70	12
19	110	48	90.5	1.88	9	71	99
20	122	56	95.7	2.09	7	75	99

Python Code (Finding VIF for all X_i)

```
import pandas as pd
import statsmodels.api as sm
from statsmodels.stats.outliers_influence import variance_inflation_factor
df = pd.read_csv("https://reneshbedre.github.io/assets/posts/reg/bp.csv")
X = df[['Age', 'Weight', 'BSA', 'Dur', 'Pulse', 'Stress']]   # independent variables
y = df['BP']   # dependent variables
X = sm.add_constant(X)
# fit the regression model
reg = sm.OLS(y, X).fit()
# get Variance Inflation Factor (VIF)
pd.DataFrame({'variables':X.columns[1:], 'VIF':[variance_inflation_factor(X.values, i+1)
for i in range(len(X.columns[1:]))]})
```

Results (VIF values)

	Variables	VIF
0	Age	1.762807
1	Weight	8.417035
2	BSA	5.328751
3	Dur	1.237309
4	Pulse	4.413575
5	Stress	1.834845

The explanatory variables (or independent variables) Weight, BSA and Pulse have somewhat large values of VIF. Lets say we now focus on these three (since all of their VIF values are above 4)

Deeper Look: The correlation matrix

- Looking at the correlation matrix can give us a deeper insight into which explanatory variables are correlated with each other

Python Code (Finding Correlation Matrix)

```
X = df[['Age', 'Weight', 'BSA', 'Dur', 'Pulse',
'Stress']]   # independent variables


# correlation analysis
X.corr()
```

Results (Correlation Matrix)

	Age	Weight	BSA	Dur	Pulse	Stress
Age	1	0.407349	0.378455	0.343792	0.618764	0.368224
Weight	0.407349	1	0.875305	0.20065	0.65934	0.034355
BSA	0.378455	0.875305	1	0.13054	0.464819	0.018446
Dur	0.343792	0.20065	0.13054	1	0.401514	0.31164
Pulse	0.618764	0.65934	0.464819	0.401514	1	0.50631
Stress	0.368224	0.034355	0.018446	0.31164	0.50631	1

Weight & BSA are strongly correlated
Weight and Pulse are also correlated
Age and Pulse are also correlated

What should be done?

- We don’t necessarily always decide to remove the explanatory variable with the highest VIF value, i.e., in this case, Weight
  - Possibly since Weight is easy to measure
  - Possibly because BSA and Pulse are comparatively harder to measure
- Let’s assume we decide to remove BSA and Pulse
  - Calculate the VIF values again after the removal

Python Code (Finding VIF for all X_i after Mod)

```
X = df[['Age', 'Weight', 'Dur', 'Stress']]   # independent variables
y = df['BP']   # dependent variables
X = sm.add_constant(X)
# Variance Inflation Factor (VIF)
pd.DataFrame({'variables':X.columns[1:],
'VIF':[variance_inflation_factor(X.values, i+1) for i in
range(len(X.columns[1:]))]})
```

Results (VIF values)

	Variables	VIF
0	Age	1.468245
1	Weight	1.234653
2	Dur	1.200060
3	Stress	1.241117

VIF values are much smaller now!

Example of managing multicollinearity

- https://www.reneshbedre.com/blog/variance-inflation-factor.html

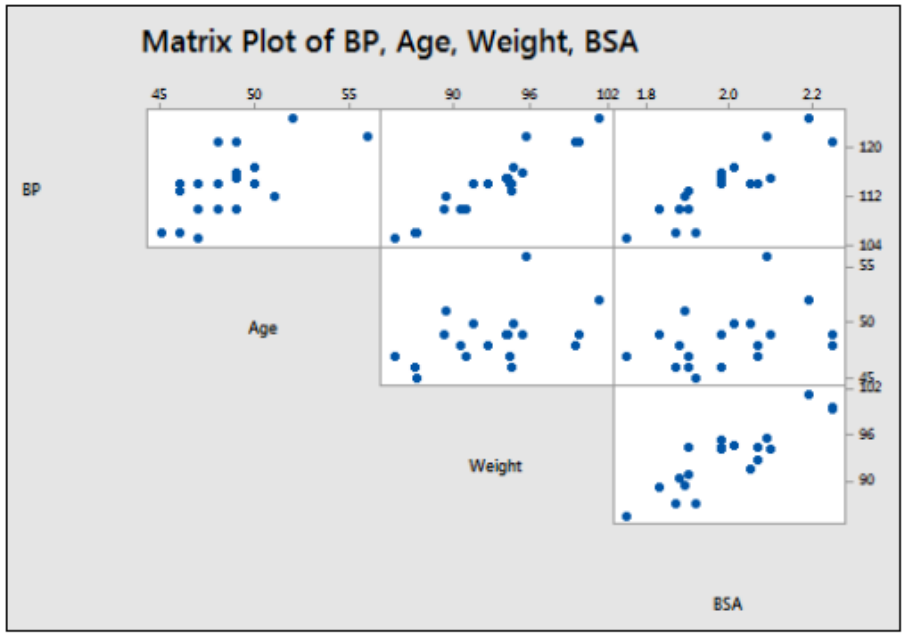
Code

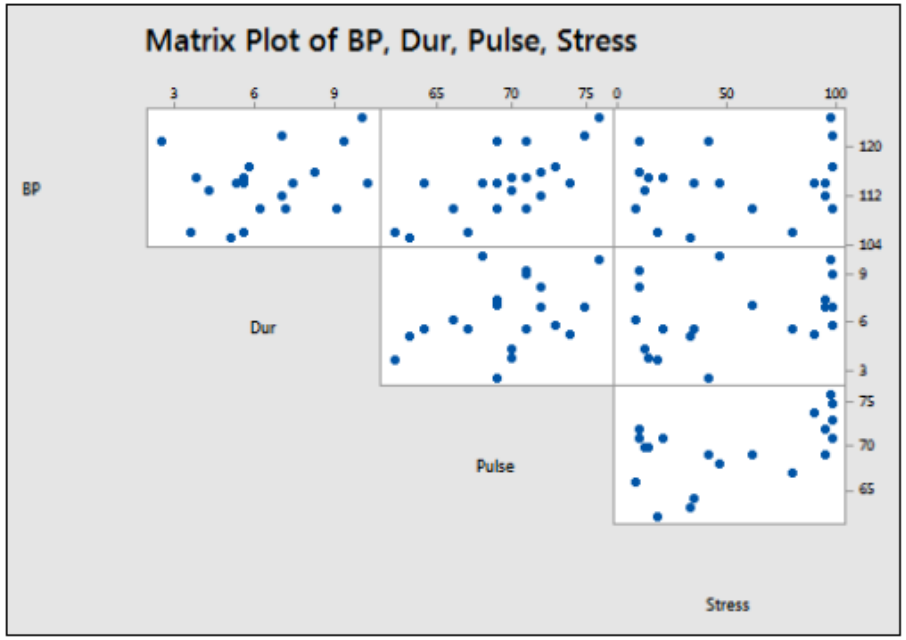
- https://colab.research.google.com/drive/1gOL8WEbL6t5C8eHZsLtdEwB1rEwH8RdB?usp=sharing

- Excel:
  - BP (CSV): https://1drv.ms/u/s!Apc0G8okxWJ1zHS32OEeh8tVtIIE?e=aT32t4
  - BP/Qty Demanded (Excel): https://1drv.ms/x/s!Apc0G8okxWJ1zHe7dOiArhibNXFd?e=yfo9Kw

By Hand

- We may find R^2 for each explanatory variable X_i by setting X_i as the response variable in place of Y, and the remaining explanatory variables as part of a separate new regression model
  - We would have to repeat this for each X_i
  - Knowing R^2 means we can easily calculate VIF


Matrix Plot of BP, Age, Weight, BSA
45
50
55
90
96
102
1.8
2.0
2.2
120
112
104
55
50
45
102
96
90
BP
Age
Weight
BSA


Matrix Plot of BP, Dur, Pulse, Stress
3
6
9
65
70
75
0
50
100
120
112
104
9
6
3
75
70
65
BP
Dur
Pulse
Stress

Useful Links & Resources

- **Source**:
  - https://online.stat.psu.edu/stat462/node/180/
- **Reference**:
  - openintro.org/os (Chapter 9, Section 9.1)