Statistical Data Mining
MATH 4720
Lecture 5
Multi-Collinearity

Gaurav Gupta

GEH A401

Multi-Collinearity

- In regression, "multicollinearity" refers to **predictors that are correlated with other predictors**.
- Multicollinearity is an issue where independent variables are not truly independent. There is **dependence structure.**
- If independent variables are highly correlated, change in one variable would cause change to another and so the model results fluctuate significantly.
- The model results will be unstable and vary a lot given a small change in the data or model.

Perfect(Exact) Multicollinearity

- If two or more independent variables have an exact linear relationship between them then we have perfect multicollinearity
- **Examples:** including the same information twice (weight in pounds and weight in kilograms), not using dummy variables correctly (falling into the dummy variable trap), etc.

Perfect(Exact) Multicollinearity…

Here is an example of perfect multicollinearity in a model with two explanatory variables:

$$Y_i = \beta_0 + \beta_1 X_{1i} + \beta_2 X_{2i} + \varepsilon_i$$

$$X_{1i} = \alpha_0 + \alpha_1 X_{2i}$$

Consequence: OLS cannot generate estimates of regression coefficients (error message)

Why? OLS cannot estimate the marginal effect of X_1 on Y while holding X_2 constant because X_2 moves exactly when X_1 moves!

Solution: Easy - Drop one of the variables!

Imperfect (or Near) Multicollinearity

When we use the word multicollinearity we are usually talking about severe imperfect multicollinearity. When explanatory variables are approximately linearly related, we have

$$Y_i = \beta_0 + \beta_1 X_{1i} + \beta_2 X_{2i} + \varepsilon_i$$

$$X_{1i} = \alpha_0 + \alpha_1 X_{2i} + u_i$$

How are correlation and collinearity different?

Collinearity is a linear association between two predictors. Multicollinearity is a situation where two or more predictors are highly linearly related. In general, an absolute correlation coefficient of >0.7 among two or more predictors indicates the presence of multicollinearity.

	OverallQual	YearBuilt	TotalBsmtSF	1stFlrSF	2ndFlrSF	GrLivArea	PoolArea	MoSold	YrSold	SalePrice
OverallQual	1	0.572323	0.537808	0.476224	0.295493	0.593007	0.0651658	0.0708152	-0.0273467	0.790982
YearBuilt	0.572323	1	0.391452	0.281986	0.0103077	0.19901	0.00494973	0.0123985	-0.0136177	0.522897
TotalBsmtSF	0.537808	0.391452	1	0.81953	-0.174512	0.454868	0.126053	0.0131962	-0.0149686	0.613581
1stFlrSF	0.476224	0.281986	0.81953	1	-0.202646	0.566024	0.131525	0.0313716	-0.0136038	0.605852
2ndFlrSF	0.295493	0.0103077	-0.174512	-0.202646	1	0.687501	0.0814869	0.0351644	-0.0286999	0.319334
GrLivArea	0.593007	0.19901	0.454868	0.566024	0.687501	1	0.170205	0.0502397	-0.0365258	0.708624
PoolArea	0.0651658	0.00494973	0.126053	0.131525	0.0814869	0.170205	1	-0.0337366	-0.0596889	0.0924035
MoSold	0.0708152	0.0123985	0.0131962	0.0313716	0.0351644	0.0502397	-0.0337366	1	-0.145721	0.0464322
YrSold	-0.0273467	-0.0136177	-0.0149686	-0.0136038	-0.0286999	-0.0365258	-0.0596889	-0.145721	1	-0.0289226
SalePrice	0.790982	0.522897	0.613581	0.605852	0.319334	0.708624	0.0924035	0.0464322	-0.0289226	1

Problems

- It would be hard for you to **choose the list of significant variables** for the model if the model gives you different results every time.
- **Coefficient Estimates would not be stable** and it would be hard for you to interpret the model.
- The unstable nature of the model may cause **overfitting**.

Diagnostics of multicollinearity

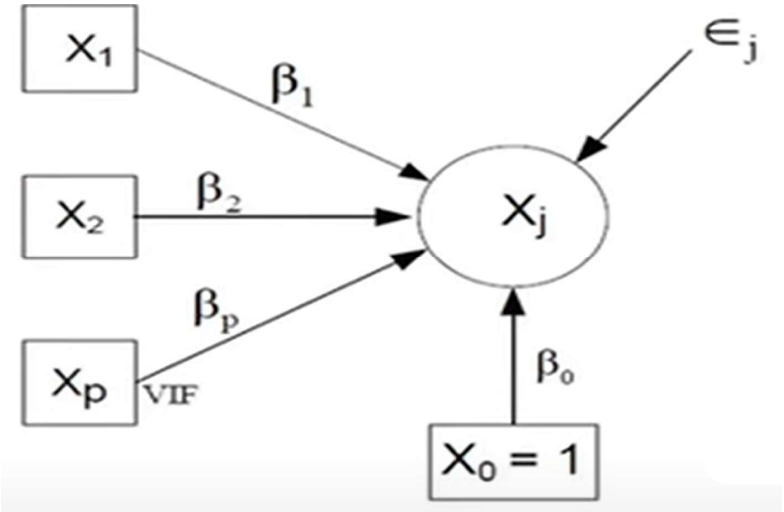
- Variance Inflation Factor (VIF) and Tolerance
- Eigen Value Structure
- Multicollinearity Condition numbers

Variance Inflation Factor (VIF) & Tolerance

The correlation coefficient indicates the strength of the linear relationship that might be existing between two variables

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

1 = not correlated.
Between 1 and 5 = moderately correlated.
Greater than 5 = highly correlated.



R_j^2	0	0.2	0.4	0.5	0.6	0.8	0.9	1
VIF	1	1.25	1.67	2	2.5	5	10	∞
1/VIF	1	0.8	0.6	0.5	0.4	0.2	0.1	0.0

Example: Blood Pressure Variance Inflation Factor Matrix

A VIF of 1 indicates two variables are not correlated, a VIF between 1 and 5 indicates moderate correlation, and a VIF above 5 indicates high correlation.

	Blood pressure	Age	Weight	Body surface area	Duration of hypertension	Puls
Age	2.93					
Weight	20.00	1.69				
Body surface area	7.46	1.61	8.00			
Duration of hypertension	1.41	1.52	1.25	1.15		
Pulse	3.58	2.62	2.93	1.87	1.67	
Stress	1.20	1.58	1.04	1.02	1.45	2.0

Eigen Value Structure and MCN

R = Correlation matrix of $X_{n \times p}$

$$= \begin{bmatrix} 1 & r_{12} & \cdots & r_{1p} \\ r_{12} & 1 & \cdots & \\ \vdots & & \ddots & \\ r_{1p} & \cdots & & 1 \end{bmatrix}_{p \times p}$$

Using Spectral Decomposition

$$R = \sum_{j=1}^{p} V_j \lambda_j V_j^T$$

λ_j = j^{th} Eigenvalue

V_j = j^{th} Eigen Vector

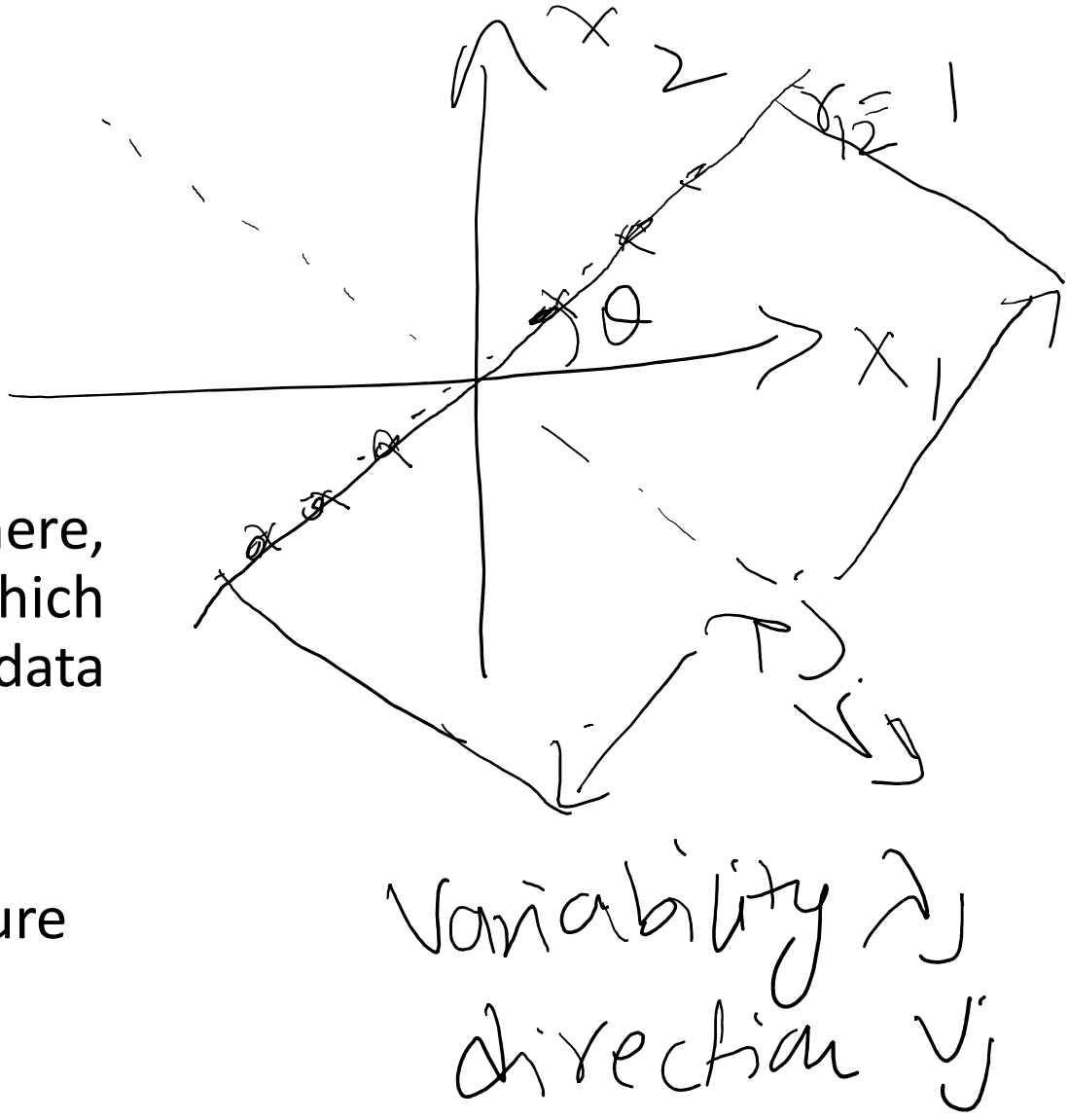
V_j: $p \times 1$, λ_j: 1×1, V_j^T: $1 \times p$

$V_j \lambda_j V_j^T$: $p \times p$

Eigenvalue

- λ Variance component
- V Direction component
- If we do some transformation here, we can get another dimension which will capture the totality of the data given here

- λ_1 and V_1 will be sufficient to capture this data



- 2 Variable: If one eigenvalue is zero then it's perfect correlation case.
- More than 2 variable:

$$\lambda_1 \geq \lambda_2 \geq \lambda_3 \ldots\ldots \geq \lambda_p$$

$$\lambda_1, \lambda_2, \lambda_3 \ldots\ldots \lambda_m \quad \lambda_{m+1} \ldots\ldots \lambda_p$$

↓

Close to zero
So there is MC.

Multi-collinearity number (MCN)

$$MCN = \frac{\lambda_1}{\lambda_p}$$

$$VIF_m \leq MCN \leq p \sum_{j=1}^{p} VIF_j$$

MCN < 100: Not Serious

MCN > 1000: very Serious

Remedies for Multicollinearity

No single solution exists that will eliminate multicollinearity. Certain approaches may be useful:

1. Do Nothing

Live with what you have.

2. Drop a Redundant Variable

If a variable is redundant, it should have never been included in the model in the first place. So dropping it actually is just correcting for a specification error. Use economic theory to guide your choice of which variable to drop.

3. Transform the Multicollinear Variables

Sometimes you can reduce multicollinearity by re-specifying the model, for instance, create a combination of the multicollinear variables. As an example, rather than including the variables GDP and population in the model, include GDP/population (GDP per capita) instead

4. Increase the Sample Size

Increasing the sample size improves the precision of an estimator and reduces the adverse effects of multicollinearity. Usually adding data though is not feasible.

Exercise:

- Check the multicollinearity of the given data. Write your conclusion based on correlation matrix and eigen value.

Note: Typically, in a small regression problem, we wouldn’t have to worry too much about collinearity. However, in cases where we are dealing with thousands of independent variables, this analysis becomes useful.