STA302 METHODS OF DATA ANALYSIS 1

MODULE 7: DECOMPOSING THE VARIANCE PART 2

PROF. KATHERINE DAIGNAULT

MODULE 7 OUTLINE

1. Decomposition & Measuring Goodness
2. Problems with Related Predictors
3. Assessing & Addressing Multicollinearity

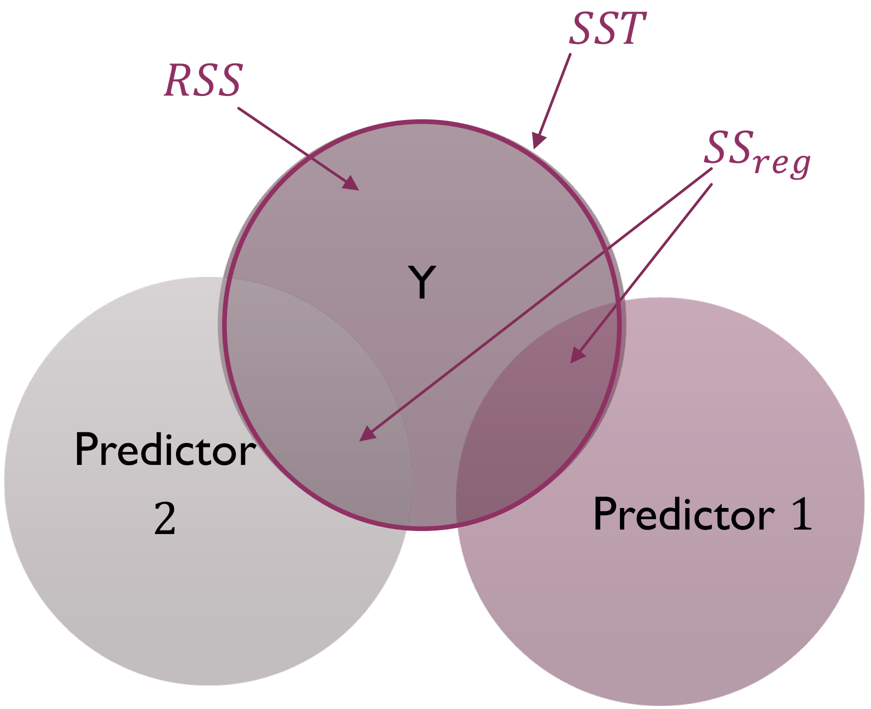
DECOMPOSITION OF SUM OF SQUARES OVERVIEW

- The decomposition of sum of squares:

  $$SST = \sum(y_i - \bar{y})^2 = \sum(\hat{y}_i - \bar{y})^2 + \sum(y_i - \hat{y}_i)^2 = SS_{reg} + RSS$$

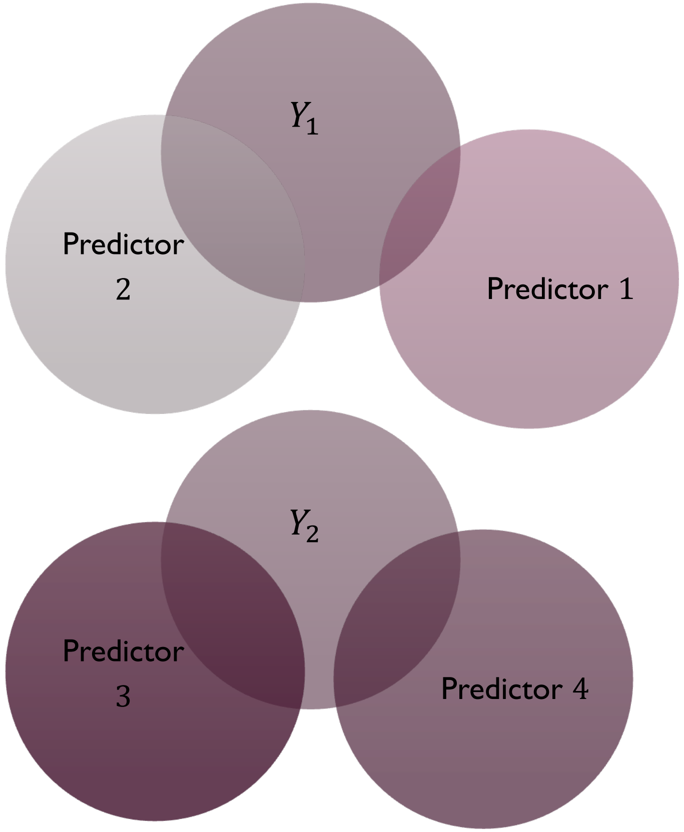
  - SST is total overall variation in response prior to fitting a model ($df = n - 1$)
  - SS_{reg} is the overall variation explained by the model ($df = p$)
  - RSS is the overall variation remaining/left unexplained by the model ($df = n - p - 1$)
- Generally, consider a good model will have $SS_{reg} > RSS$
- A model with more predictors (even non-significant ones) has smaller RSS so

  $$RSS_{small} > RSS_{big} \Leftrightarrow SS_{reg,small} < SS_{reg,big}$$



GOODNESS OF A MODEL

- Consider two statisticians working on a similar research problem.
  - they are given different dataset to work with a response variable measuring the same characteristic but in two different ways
  - each selects a 2-predictor model, but no common predictors
- Looking at each model's SS_{reg} is limited because each model's SST is different
  - similarly, the ANOVA test of significance cannot tell us which model is better
    - gives existence of linear relationship but does not quantify relative goodness
- Would need a measure that is unitless (i.e. one that doesn't depend on how the response was measured).

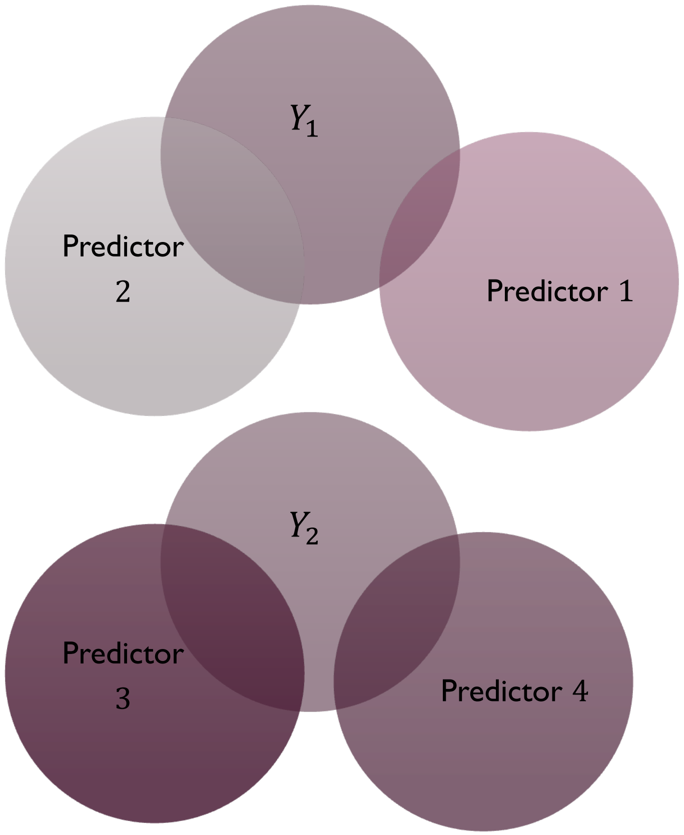


COEFFICIENT OF DETERMINATION (R^2)

- Working on different responses means a larger SS_{reg} could be a result of an overall larger SST.
- "Standardize" the variation explained by the SST so no longer dependent on starting variation
- Resulting measure of goodness is called the Coefficient of Determination (R^2), given by

$$R^2 = \frac{SS_{reg}}{SST} = 1 - \frac{RSS}{SST}$$

- In simple linear regression, $R^2 = (r)^2$, the sample correlation squared
- $0 \leq R^2 \leq 1$ so it represents the proportion of variation in the response that has been explained by the model

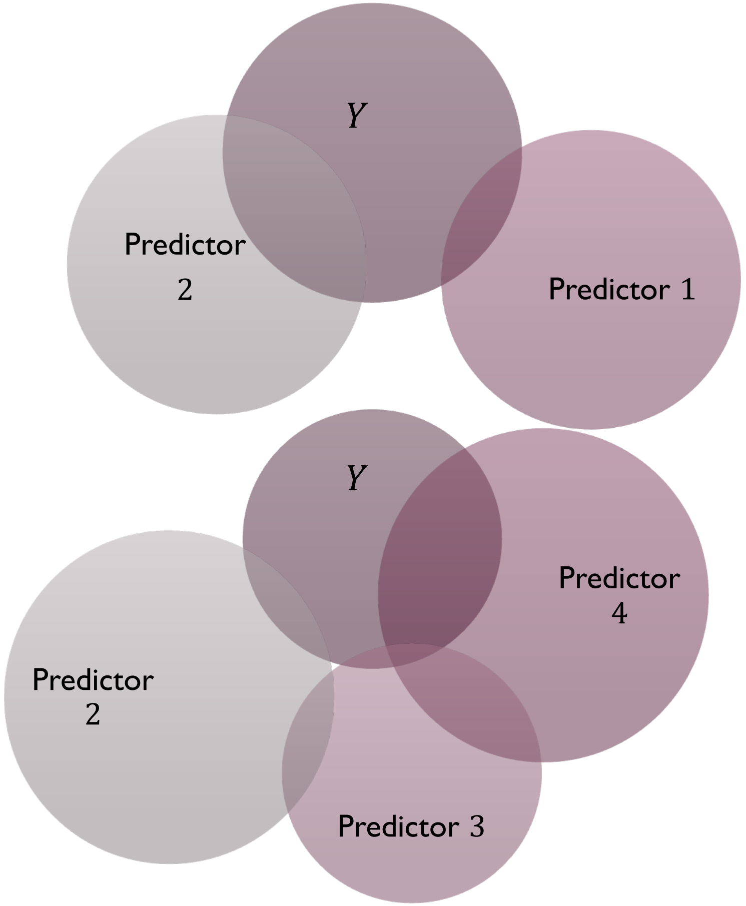


ADJUSTED COEFFICIENT OF DETERMINATION (R^2_{adj})

- When comparing models with different numbers of predictors, $SS_{reg,big} > SS_{reg,small}$
  - A bigger model will also have a larger R^2 even if extra predictors not significant
- Like the ANOVA test/Partial F test, adjust decomposition with degrees of freedom:

$$R^2_{adj} = 1 - \frac{RSS/(n-p-1)}{SST/(n-1)}$$

- Loses interpretation of "proportion of variation explained by model"
- Will say a bigger model is better only if SS_{reg} has increased enough to compensate for adding complexity to the model



EXAMPLE BY HAND & USING R

A model involving 3 predictors (X_1, X_2, X_3) is fit to a response Y using a sample of 30.

The response has sample variance $s_y^2 = 376.6853$.

The model yields an estimated error variance of $s^2 = 50.555$.

Compute the two coefficients of determination for this model.

1. Collect given values:

$$n = 30, \qquad p = 3,$$
$$s_y^2 = {SST}/{(n-1)} = 376.6853$$
$$s^2 = {RSS}/{(n-p-1)} = 50.555$$

2. Find decomposition pieces:

$$SST = (n-1)s_y^2 = 29(376.6853) = 10923.87$$

$$RSS = (n-p-1)s^2 = 26(50.555) = 1314.43$$

3. Compute Coefficients:

88% of variation in Y explained by model

$$R^2 = 1 - \frac{RSS}{SST} = 1 - \frac{1314.43}{10923.87} \approx 0.88$$

$$R_{adj}^2 = 1 - \frac{RSS/(n-p-1)}{SST/(n-1)} = 1 - \frac{50.555}{376.6853} \approx 0.866$$

smaller than R^2

```
> def <- read.table("defects.txt", header=T)
> model <- lm(Defective ~ Temperature + Density + Rate, data=def)
> summary(model)

Call:
lm(formula = Defective ~ Temperature + Density + Rate, data = def)

Residuals:
     Min       1Q   Median       3Q      Max 
-12.7367  -4.1116  -0.5755   2.7617  16.3279 

Coefficients:
            Estimate Std. Error t value Pr(>|t|)  
(Intercept)  10.3244    65.9265   0.157   0.8768  
Temperature  16.0779     8.2941   1.938   0.0635 .
Density      -1.8273     1.4971  -1.221   0.2332  
Rate          0.1167     0.1306   0.894   0.3797  
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 7.11 on 26 degrees of freedom
Multiple R-squared:  0.8797,	Adjusted R-squared:  0.8658 
F-statistic: 63.36 on 3 and 26 DF,  p-value: 4.371e-12
```

MODULE 6 OUTLINE

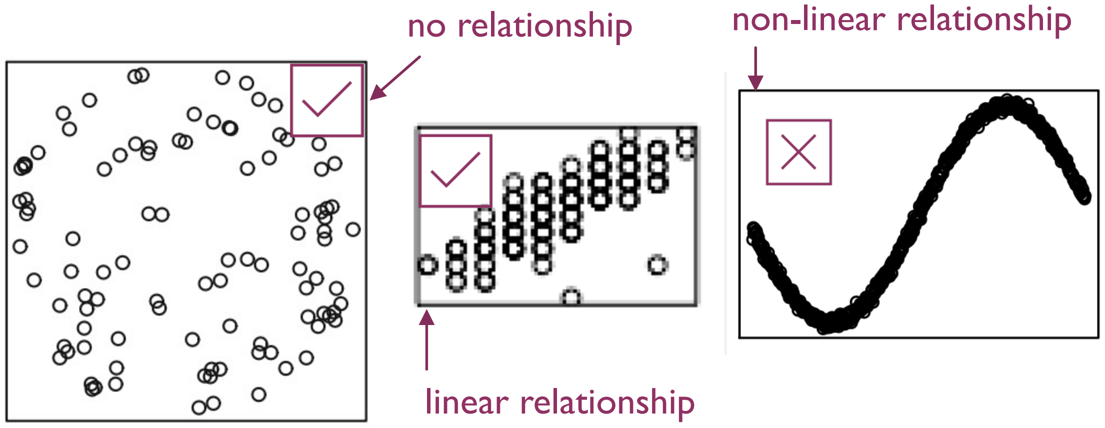
1. Decomposition & Measuring Goodness
2. Problems with Related Predictors
3. Assessing & Addressing Multicollinearity

RELATED PREDICTORS & CONDITION 2

- **Conditional mean predictor condition:** the mean of each predictor is related to each other predictor in no more complicated way than linearly

$$E(X_i|X_j) = \alpha_0 + \alpha_1 X_j$$

  - Linear or no relationship satisfy condition; anything else violates



- Ideally, want to observe no relationship between predictors
- A linear trend between predictors means predictors are correlated or collinear (i.e., a linear relationship exists between them)
- Correlation measures strength of <u>linear</u> association between <u>two</u> continuous variables.
  - if small (between 0 and 0.4, or negative version), we see little to no relationship
  - if moderate (between 0.4 to 0.6 , or negative version), a more obvious linear trend appears
  - if strong (between 0.6 and 1 , or negative version), a clear linear trend appears

RANK OF THE DESIGN MATRIX & CORRELATION

- Perfect correlation is particularly of concern
  - occurs when we have sample correlations of +1 or -1
- Means that one predictor is perfectly linearly related to another
- In regression, this corresponds to having columns of our X matrix be functions of one another
  - When this happens, we say that our matrix is not full column rank (i.e., we have dependent/related columns)
- This results in having issues finding the inverse $(\boldsymbol{X}^T\boldsymbol{X})^{-1}$ so we can't estimate the linear relationship

Example:

$$\begin{pmatrix} 1 & 4 & 3 \\ 2 & 2 & 3 \\ 3 & 8 & 7 \end{pmatrix}$$

We can find an equation that allows one column to be written as a function of the others:

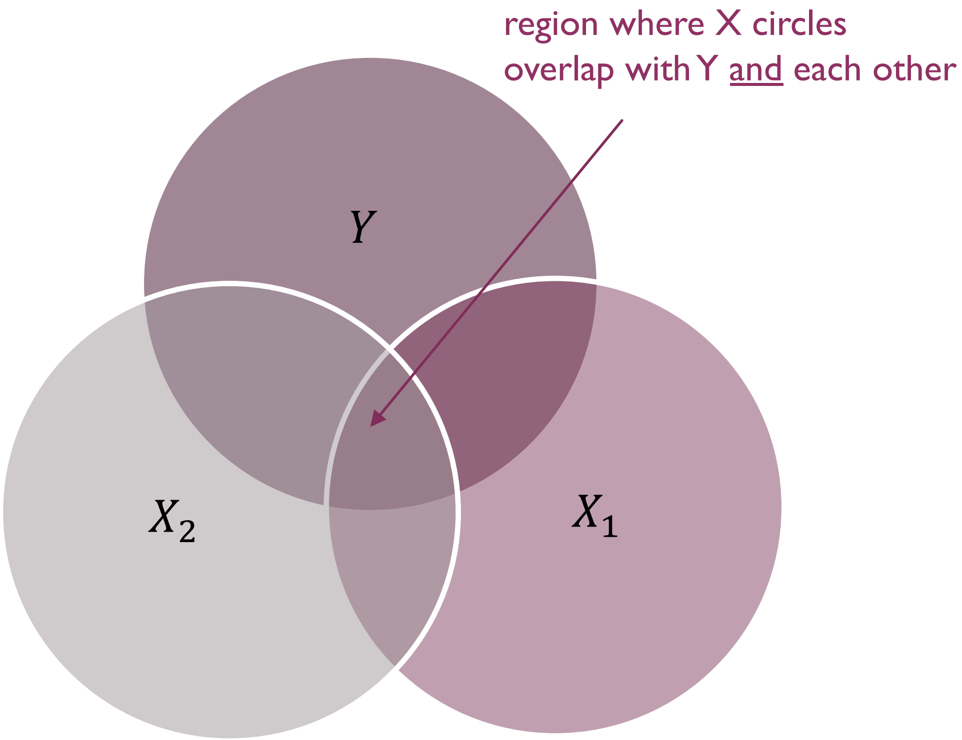
Column 3 = Column 1 +0.5(Column 2)

Check that this is the case:

- 3 = 1+0.5*4 = 1+2
- 3 = 2+0.5*2 = 2+1
- 7 = 3 + 0.5*8 = 3+4

RANK OF DESIGN MATRIX & MULTICOLLINEARITY

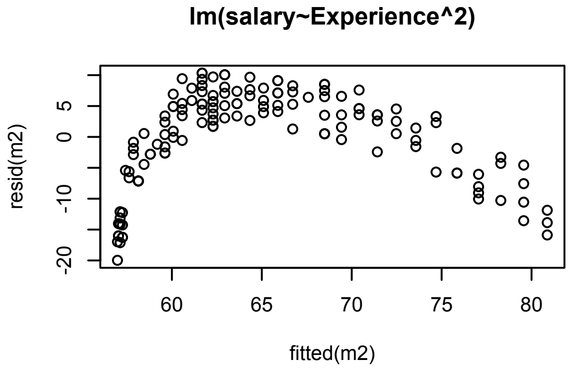
- Correlation only tells you if any 2 predictors are related
- We want to know if possibly more than 2 predictors are related
  - this is called multicollinearity
- The strength of the impact of this relationship is due not only to how related the predictors are
  - Also how much those predictors explain Y to begin with (i.e., the overlap with Y)
- Creates an instability in how the predictors are used to estimate the mean responses
  - the model can't distinguish between how much variation is due only to X_1 versus only to X_2



IMPACT OF CORRELATED PREDICTORS

- Geometrically, regression is finding a vector perpendicular to a model plane built of X's
  - Each X creates a support that defines this plane and makes it stable
  - Multicollinearity makes the surface unstable so it wobbles/collapses like a poorly constructed deck
- This causes many potential problems:
  - wrong estimated coefficients: coefficients may have wrong sign compared to literature
  - contradictory significance: many predictors may be insignificant when overall F test is highly significant
  - inflated variances: standard errors of estimated coefficients are much larger than they should be

http://laurenhorner.blogspot.com/2014/04/its-deck-saster.html



STRUCTURAL MULTICOLLINEARITY

- Multicollinearity can cause problems in properly understanding the relationship
  - often this is out of our control and arises due to how variables defined or related in reality
- However, there are times where we create multicollinearity on purpose:
  - Plots show how we arrive at fitting a polynomial regression model
    - the model with best residual plot is $\hat{y}_i = \hat{\beta}_0 + \hat{\beta}_1 x_i + \hat{\beta}_2 x_i^2$
    - means that column 3 of design matrix is $(\text{column } 2)^2$
  - Using an interaction term and a main effect term in a model
    - we intentionally create a predictor (the interaction) that is a function of the two predictors being interacted
- These situations should be noted but are not of great concern as the structure we create is needed usually to satisfy assumptions

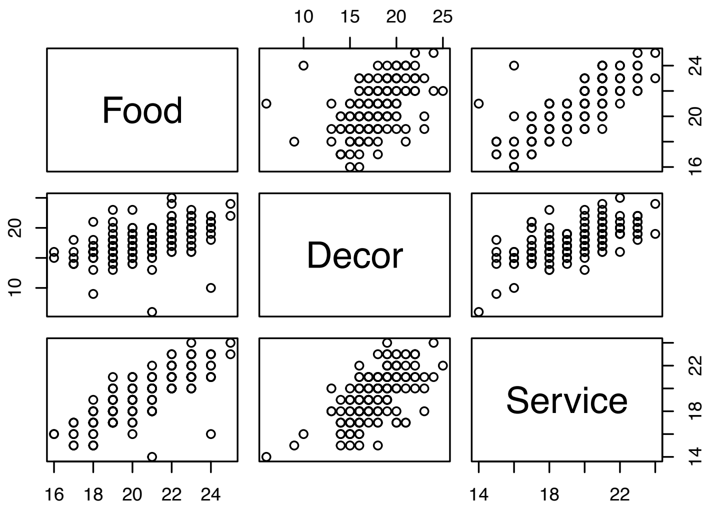
MODULE 6 OUTLINE

1. Decomposition & Measuring Goodness
2. Problems with Related Predictors
3. Assessing & Addressing Multicollinearity

PRELIMINARY ASSESSMENT IN DATA EXPLORATION

- Preliminary assessment can help in justifying decisions to handle multicollinearity
- Look at pairwise correlations among predictors:
  - very limited since only two predictors at a time
  - know early (in EDA) but doesn't convey extent of problem
  - can be used accidently on non-linear relationships
- Look at condition 2 assessment plots:
  - limited since only visualize two predictors at a time
  - allows identification of non-linear relationship that are not a multicollinearity issue (but are a condition 2 issue)
- Neither considers conditionality of predictors with each other and with the response

```
> cor(nyc[,c(4:6)])
             Food     Decor   Service
Food    1.0000000 0.5039161 0.7945248
Decor   0.5039161 1.0000000 0.6453306
Service 0.7945248 0.6453306 1.0000000
```



VARIANCE INFLATION FACTOR (2 PREDICTOR CASE)

- To formally check if multicollinearity is present, we compute a measure called the variance inflation factor (VIF)
- It quantifies how much larger the variance of a coefficient is due to multicollinearity.
- For simplicity, consider a two-predictor model:

$$y_i = \beta_0 + \beta_1 x_{i1} + \beta_2 x_{i2} + \varepsilon_i$$

- We define the following values:
  - r_{12} is the sample correlation between x_1 and x_2
  - s_{x_j} is the sample standard deviation for x_j

- It can be shown that $Var(\hat{\beta}_j)$ is written as

$$Var(\hat{\beta}_j) = \frac{1}{1-(r_{12})^2} \times \frac{\sigma^2}{(n-1)s_{x_j}^2}, \qquad j = 1,2$$

- The first term is the VIF, measuring how much larger the variance is as a result of stronger correlation
  - As r_{12} grows towards $+1$ or decreases towards -1:
    - r_{12}^2 also grows towards 1
    - denominator shrinks towards 0
    - VIF increases in value indicating variance has inflated

VARIANCE INFLATION FACTOR (2+ PREDICTOR CASE)

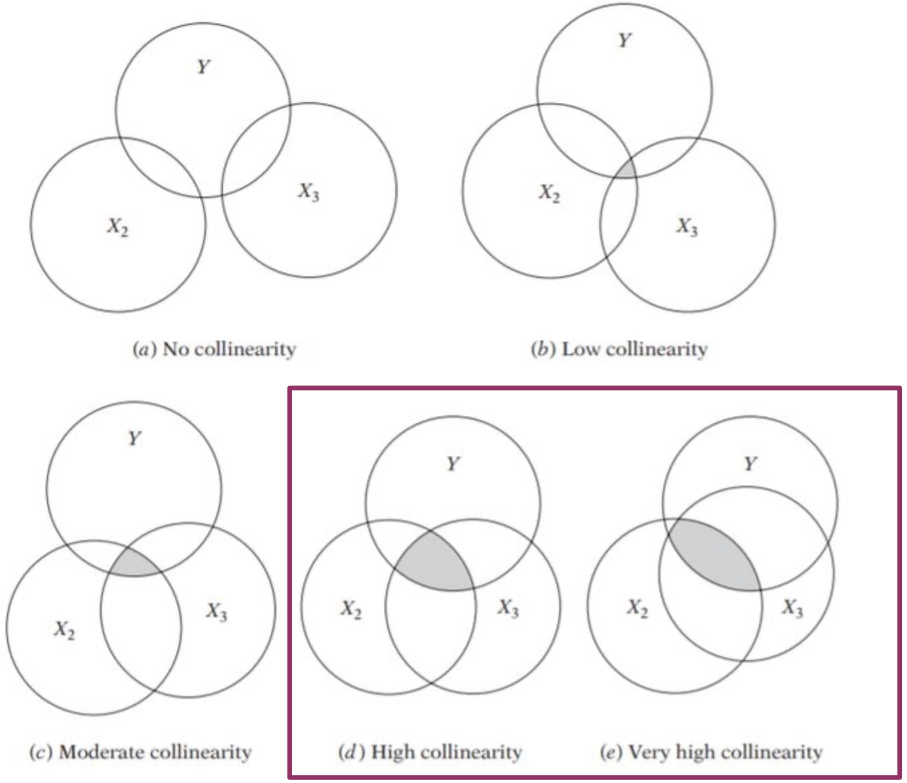
- Case of more than two predictors requires a replacement for r_{12}
  - need a measure that also quantifies strength or goodness of linear relationship
  - saw that R^2 measures this as proportion of variance explained by linear relationship
- Here R^2 would be used to say that a model explains variation in a predictor, not a response
  - e.g. suppose we have 3 predictors in our model
  - using, e.g., X_1 as a response, can create a model of the form $X_1 = \beta_0 + \beta_1 x_2 + \beta_2 x_3 + \varepsilon$
  - R^2 of this model measures strength of linear relationship between X_1 and both X_2 and X_3
- R^2 will be used like r_{12} to say how the variance is inflated because a strong linear relationship exists between predictors
- A similar expression for the variance of any $\hat{\beta}_j$ is

$$Var(\hat{\beta}_j) = \frac{1}{1 - R_j^2} \times \frac{\sigma^2}{(n-1)s_{x_j}^2}, \qquad j = 1, \dots, p$$

- The VIF for $\hat{\beta}_j$ is again the first term, incorporating the R^2 from a model using X_j as response
- Note: unlike the correlation that was squared, R_j^2 is not being squared
  - the square is part of the notation for the quantity

VARIANCE INFLATION FACTOR CONCLUSIONS

- The VIF form we use is $\frac{1}{1-R_j^2}$ as it generalizes to models of any size, but only calculable for numerical predictors.
- The stronger the proportion of variation in X_j explained by the remaining $p-1$ predictors, the larger the inflation
- We generally use a cutoff of $VIF > 5$ to identify severe multicollinearity
  - technically any $VIF > 1$ means some multicollinearity present
  - means variances are very inflated which could lead to incorrect conclusions regarding significance
- Note, reliability of conclusion depends on assumptions of model holding



https://www.researchgate.net/publication/348558181_Multicollinearity/figures?lo=1

EXAMPLE BY HAND & USING R

A model involving 3 predictors (X_1, X_2, X_3) is fit to a response Y using a sample of 30.

A model fit using X_1 as response and X_2 and X_3 as predictors yields an $R^2 = 0.9255$.

Find the VIF for X_1.

$$VIF = \frac{1}{1 - R_1^2}$$
$$= \frac{1}{1 - 0.9255} = 13.42$$

```
> summary(lm(Temperature ~ Density+Rate, data=def))

Call:
lm(formula = Temperature ~ Density + Rate, data = def)

Residuals:
     Min       1Q   Median       3Q      Max 
-0.27411 -0.07731 -0.00463  0.08016  0.40252 

Coefficients:
             Estimate Std. Error t value Pr(>|t|)    
(Intercept)  4.671889   1.237584   3.775   0.0008 ***
Density     -0.136819   0.022657  -6.039 1.91e-06 ***
Rate         0.004190   0.002922   1.434   0.1630    
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 0.165 on 27 degrees of freedom
Multiple R-squared:  0.9255,	Adjusted R-squared:  0.92 
F-statistic: 167.8 on 2 and 27 DF,  p-value: 5.892e-16
```

~93% of variation in X_1 explained by other two predictors

```
> 1/(1-0.9255)
[1] 13.42282
```

variance of $\hat{\beta}_1$ inflated by factor of >13

```
> library(car)
Loading required package: carData
> model <- lm(Defective ~ Temperature + Density + Rate, data=def)
> vif(model)
Temperature     Density        Rate 
  13.431614   14.508872    6.642619 
```

variance of $\hat{\beta}_1$ inflated by factor of >13

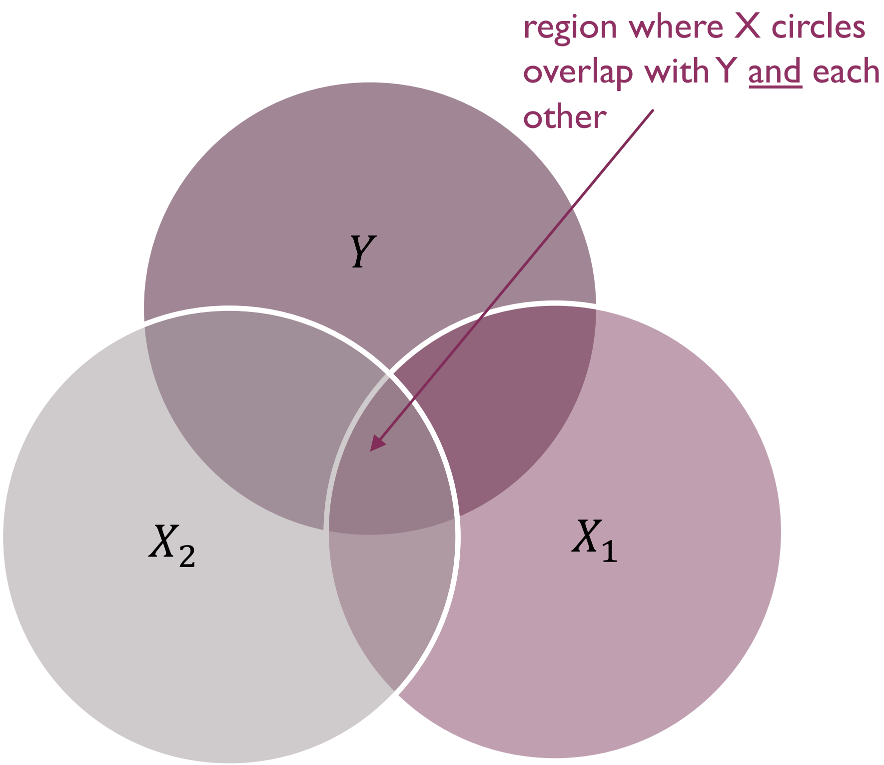
variance of $\hat{\beta}_2$ inflated by factor of >14

variance of $\hat{\beta}_2$ inflated by factor of >6

Overall, we have serious multicollinearity present between all predictors in this model

ADDRESSING MULTICOLLINEARITY

- Really only two main ways to address severe multicollinearity
  - collect more data in the hopes of seeing less correlation due to more variation in the data → not feasible/practice
  - respecify the model (i.e. remove /change the form of some predictors) → most popular
- Multicollinearity ⟹ related predictors explain identical parts of SST
  - removing at least one removes some of the common overlap pictured
- Care should be taken in the choice of predictors to remove:
  - don't generally remove a predictor of interest to your research question
  - consider how choice impacts model goodness (e.g. assumptions, prediction, overall fit, etc.)



MODULE TAKE-AWAYS

1. How are both coefficients of determination computed?
2. What is the difference between the two coefficients and why do we need two?
3. How are the coefficients of determination interpreted?
4. What does the VIF measure and why is that useful for detecting multicollinearity?
5. How do we identify multicollinearity and why is it important to identify?