Regression Cautions

Stat 230: Applied Regression Analysis

Multicollinearity

Situation where 2+ predictors are “nearly” perfectly correlated

Think about it

- Work through Examples 1-3 on worksheet
- Work with your group to complete the tasks
- Let me know when finish Task 6
- Be prepared to share thoughts with the class

Diagnosing multicollinearity

- Examine scatterplot matrix
- Examine pairwise correlations between predictors
- Look for $\widehat{\beta}_i$s with unusual signs
- Notice sensitivity to changes in the model/order of predictors
- Calculate the **Variance Inflation Factor (VIF)**

Variance Inflation Factor (VIF)

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

$R_i^2 = R^2$ from model predicting x_i using all other predictor variables

Guidelines:

- $\text{VIF}_i > 5$ suspicions begin; $R_i^2 > .8$
- $\text{VIF}_i > 10$ indicates a problem; $R_i^2 > .9$
- $\text{VIF}_i > 100$ indicates a big problem; $R_i^2 > .99$

VIFs for Example 1

```r
ex1_mod <- lm(y ~ x1 + x2, data = ex1)
car::vif(ex1_mod)
```

Variable	VIF
x1	1
x2	1

VIFs for Example 2

`car::vif()` throws an error! Why?

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

- Since x_1 and x_2 are perfectly correlated, $R_i^2 = 1$
- This makes the denominator $1 - R_i^2 = 0$
- So, VIF $= \frac{1}{0}$ is undefined

VIFs for Example 3

```
1 bodyfat_mod <- lm(body_fat ~ ., data = bodyfat)
2 car::vif(bodyfat_mod)
```

Variable	VIF
triceps_skinfold	708.8
thigh_circumference	564.3
midarm_circumference	104.6

Remedial measures

- Use the model only for prediction, if you think future observations will have the same relationships amongst variables
- For polynomials, use `poly(x, degree)` in R, or center the variable
- Drop some of highly correlated variables — USE CAUTION, doesn't always work and loses information
- Add cases that break the correlation between predictors – reasonable in designed experiments
- Create composite variables (e.g., sums, averages, principal components) – can be harder to interpret
- Use ridge regression or LASSO – need Stat 270 (Statistical Learning)

The problem with R^2

Different polynomial models were used to predict the horizontal distance based on the starting point in Galileo's experimental data

Order	R^2
1	0.9264
2	0.9903
3	0.9994
4	0.9998
5	0.99996
6	1.0000

Adjusted R^2

Imposes a **penalty for model complexity**, so it increases only if the improvement outweighs the cost of making the model more complex

$$R^2_{\text{adj}} = 1 - \left(\frac{n-1}{n-p-1} \right) \cdot \left(1 - R^2\right)$$

Adjusted R^2

Order	R^2	R^2_{adj}
1	0.9264	0.91164
2	0.9903	0.98551
3	0.9994	0.99875
4	0.9998	0.9995
5	0.99996	0.99977
6	1.0000	undefined