IDS 702

Multicollinearity

Multicollinearity: the problem

- You cannot include two variables with a perfect linear association as predictors in regression

```
Coefficients: (1 not defined because of singularities)
            Estimate Std. Error t value Pr(>|t|)
(Intercept)  0.04536    0.11495   0.395    0.694
X1           4.07300    0.11732  34.716   <2e-16 ***
X2                NA         NA      NA       NA
```

Multicollinearity

- In real data, when predictors are collinear, we see standard errors inflate (which is bad)
- When might we get close:
  - Very high correlations ($|\rho| > 0.9$) among two or more predictors
  - When one or more variables are nearly a linear combination of others

Identifying multicollinearity

- Think about it during EDA
  - Is one variable derived from another?
  - Do you expect a variable to always increase as another increases?
- Look at a correlation matrix of predictors
- Look at Variance Inflation Factor (VIF): measures how much the multicollinearity between a variable and other variables inflates the variance of the regression coefficient for that variable

VIF

- $VIF_j = \frac{1}{1 - R^2_{X_j|X_{-j}}}$
- VIF will always be $\geq$1 (Why?)
- Generally, VIF=
  - 1 $\Longrightarrow$not correlated
  - Between 1 and 5 $\Longrightarrow$moderately correlated
  - Greater than 5 $\Longrightarrow$highly correlated
  - Greater than 10 $\Longrightarrow$HIGHLY correlated and we want to do something about it

VIF in R

- vif(model) gives the VIF value for each predictor
- vif() function is in the car package

```
> auto_mod <- lm(mpg~displacement+horsepower+weight+
acceleration,data=Auto)
> vif(auto_mod)
displacement   horsepower       weight
   10.686922     8.823022    10.284456
acceleration
    2.603255
```

```
> auto_mod <- lm(mpg~displacement+horsepower+weight+
acceleration+factor(origin),data=Auto)
> vif(auto_mod)
                     GVIF Df GVIF^(1/(2*Df))
displacement    13.022819  1        3.608714
horsepower       9.378330  1        3.062406
weight          10.492346  1        3.239189
acceleration     2.604609  1        1.613880
factor(origin)   1.985527  2        1.187050
```

What to do?

- Can remove one of the predictors
- Can scale the variables
- Tends to be unimportant in large samples
- Think through the application!