STAT021 Statistical Methods II

Lecture 19 Multicollinearity and Model Selection

Lu Chen
Swarthmore College
11/15/2018

Outline

Multicollinearity

- Definition and examples
- Detect multicollinearity
  - Scatterplot matrix
  - Correlation matrix
  - Variance inflation factor (VIF)
- Deal with multicollinearity

Model selection criteria

- Nested F test
- Adjusted R^2
- Mallow's C_p
- Akaike/An information criterion (AIC)

Perch model comparisons

Weight ~	F	R^2	R^2_{adj}
5. $L + W$	396.1	0.9373	0.9349
6. $L + L^2 + W$	1114	0.9847	0.9838
7. $L + W + W^2$	927	0.9816	0.9806
8. $L + L^2 + W + W^2$	865.5	0.9855	0.9843
9. $L + W + LW$	1115	0.9847	0.9838
10. $L + L^2 + W + LW$	840.9	0.9851	0.9839
11. $L + W + W^2 + LW$	820	0.9847	0.9835
12. $L + L^2 + W + W^2 + LW$	704.6	0.9860	0.9846

F tests of all the eight models have $P < 2.2 \times 10^{-16}$.

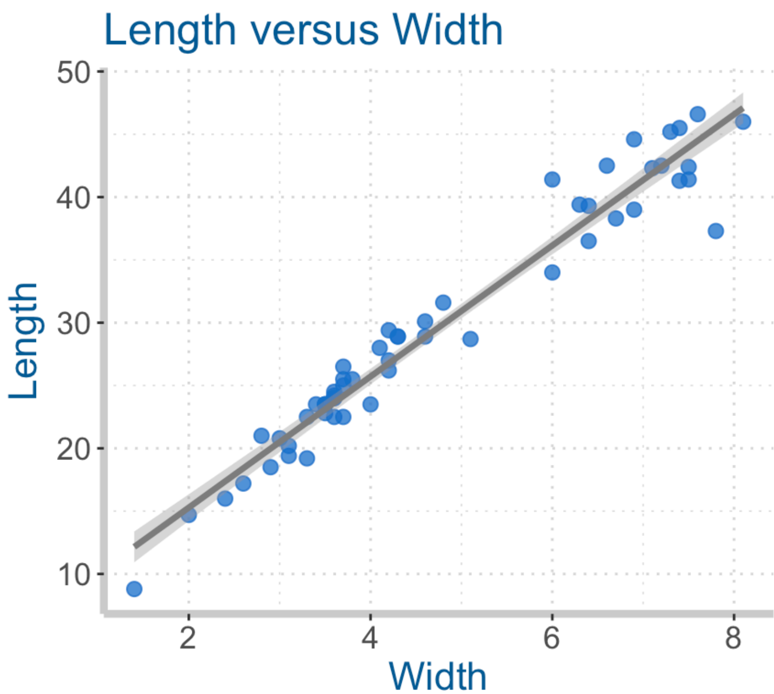
- The complete second-order model (model 12) has the largest R^2 and adjusted R^2.
- Model 12 is not significantly better than model 9 or model 6.
- Model 6 and 9 are very similar. It seems to suggest that the quadratic term L^2 and the interaction term LW have similar effect in the model.

Multicollinearity

> A set of predictors exhibits **multicollinearity** when one or more of the predictors is **strongly** correlated with some combination of the other predictors in the set.

- Multicollinearity: predictors strongly correlated with each other.
- It is NOT necessarily a "bad" thing.
  - If the predictors are related to the response variable, it is not surprising that they are related to each other.
- However, strong correlation between predictors may lead to difficulty in *fitting* and *interpreting* the model.

Multicollinearity - Example



- *Length* and *Width* are strongly correlated.
- Knowing one allows us to know the other almost for sure.
- When both are predictors for *Weight*, it is hard for the software to search for the estimates of the intercept and slopes.
- Extreme case: two predictors are perfectly correlated.

```
cor(perch$Length, perch$Width)
```

```
## [1] 0.9751074
```

Multicollinearity - Example

```r
Z <- perch$Length # A new variable Z is exactly the same as Length
cor(perch$Length, Z) # correlation of Length and Z is 1
```

```
## [1] 1
```

```r
summary(lm(Weight ~ Length + Z, data=perch))
```

```
##
## Coefficients: (1 not defined because of singularities)
##             Estimate Std. Error t value Pr(>|t|)
## (Intercept) -652.787     43.407  -15.04   <2e-16 ***
## Length        35.001      1.398   25.03   <2e-16 ***
## Z                 NA         NA      NA       NA
##
## Residual standard error: 98.82 on 54 degrees of freedom
## Multiple R-squared:  0.9207, Adjusted R-squared:  0.9192
## F-statistic: 626.5 on 1 and 54 DF,  p-value: < 2.2e-16
```

- When both *Length* and *Z* are in a regression model, R could not distinguish the two and find the slopes for both predictors. In this case, R removes one of them and fits the model.

Multicollinearity - Example

```
Z[1] <- 12 # change the first observation of Z to 12
perch$Length
```

```
##  [1]  8.8 14.7 16.0 17.2 18.5 19.2 19.4 20.2 20.8 21.0 22.5 22.5 22.5
## [14] 22.8 23.5 23.5 23.5 23.5 23.5 24.0 24.0 24.2 24.5 25.0 25.5 25.5
## [27] 26.2 26.5 27.0 28.0 28.7 28.9 28.9 28.9 29.4 30.1 31.6 34.0 36.5
## [40] 37.3 39.0 38.3 39.4 39.3 41.4 41.4 41.3 42.3 42.5 42.4 42.5 44.6
## [53] 45.2 45.5 46.0 46.6
```

```
Z
```

```
##  [1] 12.0 14.7 16.0 17.2 18.5 19.2 19.4 20.2 20.8 21.0 22.5 22.5 22.5
## [14] 22.8 23.5 23.5 23.5 23.5 23.5 24.0 24.0 24.2 24.5 25.0 25.5 25.5
## [27] 26.2 26.5 27.0 28.0 28.7 28.9 28.9 28.9 29.4 30.1 31.6 34.0 36.5
## [40] 37.3 39.0 38.3 39.4 39.3 41.4 41.4 41.3 42.3 42.5 42.4 42.5 44.6
## [53] 45.2 45.5 46.0 46.6
```

```
cor(perch$Length, Z)
```

```
## [1] 0.9990582
```

Multicollinearity - Example

```r
summary(lm(Weight ~ Length + Z, data=perch)) # Note: this is a new Z
```

```
## Coefficients:
##             Estimate Std. Error t value Pr(>|t|)
## (Intercept)  -707.92      39.79 -17.793  < 2e-16 ***
## Length        -85.71      27.98  -3.064  0.00343 **
## Z             122.34      28.33   4.319 6.91e-05 ***
##
## Residual standard error: 85.79 on 53 degrees of freedom
## Multiple R-squared:  0.9413, Adjusted R-squared:  0.9391
## F-statistic:   425 on 2 and 53 DF,  p-value: < 2.2e-16
```

- Model without Z: $\widehat{Weight} = -652.8 + 35.0 \times L(P < 2 \times 10^{-16})$
- Model with the new Z: $\widehat{Weight} = -707.9 - 85.7 \times L + 122.3 \times Z$
- *Length* becomes less significant and the relationship between *Weight* and *Length* becomes hard to interpret.

Multicollinearity - Example

```
summary(lm(Weight ~ Length, data=perch))$coefficients
```

```
## Coefficients:
##             Estimate Std. Error t value Pr(>|t|)
## (Intercept) -652.787     43.407  -15.04   <2e-16 ***
## Length        35.001      1.398   25.03   <2e-16 ***
```

```
summary(lm(Weight ~ Length + Width, data=perch))$coefficients
```

```
## Coefficients:
##             Estimate Std. Error t value Pr(>|t|)
## (Intercept) -578.758     43.667 -13.254  < 2e-16 ***
## Length        14.307      5.659   2.528 0.014475 *
## Width        113.500     30.265   3.750 0.000439 ***
```

- The model for *Weight* ~ *Length* + *Width* has similar problem.
- *Length* and *Width* provide very similar information in explaining the variability of *Weight* - redundant information makes it hard to estimate and interpret the model.

Multicollinearity - Example

```r
summary(m1 <- lm(Weight ~ Length, data=perch))$coefficients
```

```
## Coefficients:
##             Estimate Std. Error t value Pr(>|t|)
## (Intercept) -652.787     43.407  -15.04   <2e-16 ***
## Length        35.001      1.398   25.03   <2e-16 ***
```

- $SE_{b_1} = 1.398$

```r
summary(m2 <- lm(Weight ~ Length + Width, data=perch))$coefficients
```

```
## Coefficients:
##             Estimate Std. Error t value Pr(>|t|)
## (Intercept) -578.758     43.667 -13.254  < 2e-16 ***
## Length        14.307      5.659   2.528 0.014475 *
## Width        113.500     30.265   3.750 0.000439 ***
```

- $SE_{b_1} = 5.659$

```r
summary(m3 <- lm(Weight ~ Length + Z, data=perch))$coefficients
```

```
## Coefficients:
##             Estimate Std. Error t value Pr(>|t|)
## (Intercept)  -707.92      39.79 -17.793  < 2e-16 ***
## Length        -85.71      27.98  -3.064  0.00343 **
## Z             122.34      28.33   4.319 6.91e-05 ***
```

- $SE_{b_1} = 27.98$

Detect multicollinearity

- The **key problem** of multicollinearity is **inflated variance** of the slope estimates.
  - When *Width* or Z is added to the model, SE of the slope for *Length* becomes much larger suggesting more uncertainty in model estimation.
  - The stronger correlation between predictors, the heavier inflation in variance.

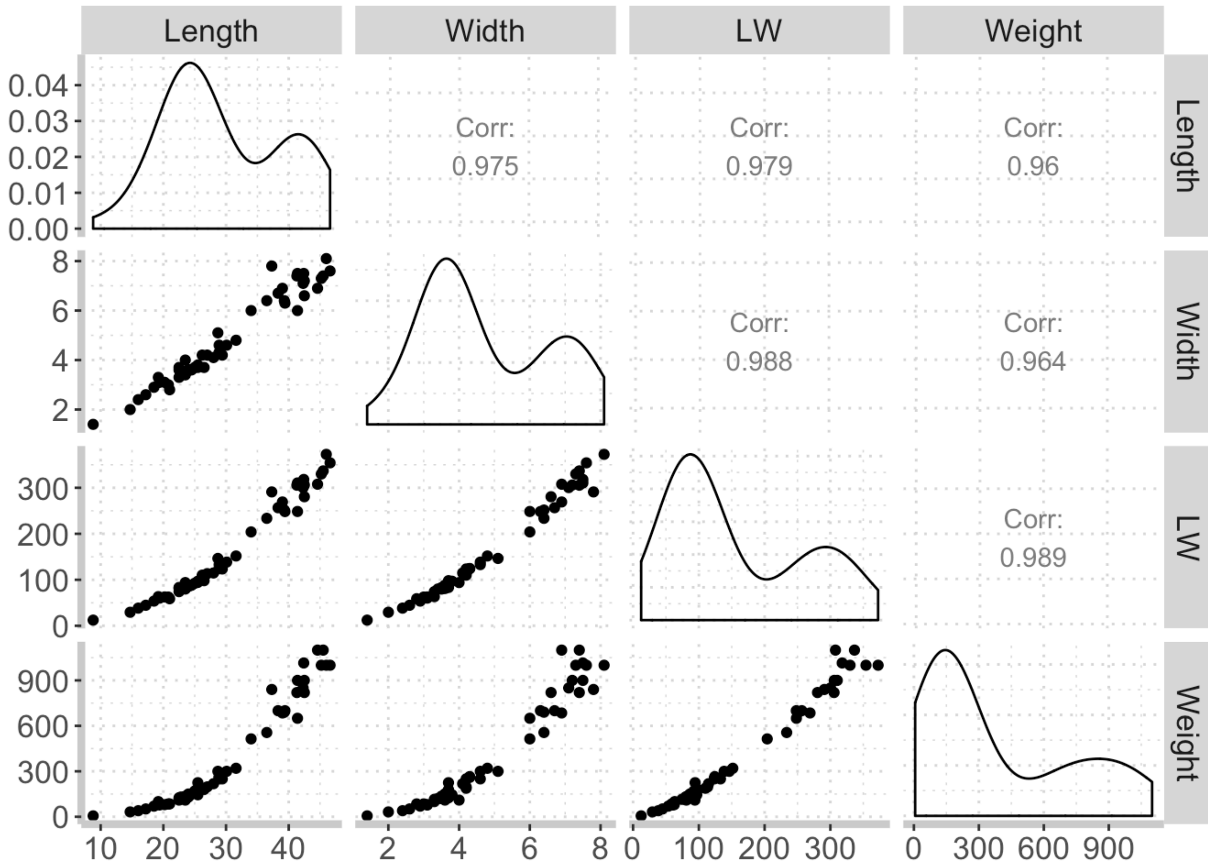
Three methods to detect multicollinearity

- Scatterplot matrix
- Correlation matrix
- Variance inflation factor

Example: $Weight \sim L + W + LW$

Scatterplot and correlation matrix

```r
perch <- data.frame(perch, LW=perch$Length*perch$Width)
library(GGally)
ggpairs(data=perch[, c("Length","Width","LW","Weight")])
```



This plot displays

1. Histograms of all variables
2. Scatterplot of any two variables
3. Correlation of any two variables

- The predictors are strongly correlated with each other.

Variance inflation factor (VIF)

For any predictor X_i in a model, the **variance inflation factor (VIF)** is computed as

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

where R_i^2 is the coefficient of multiple determination for a model to predict X_i using the other predictors in the model.

As a rough rule, we suspect multicollinearity with predictors for which the $VIF > 5$, which is equivalent to $R_i^2 > 80\%$.

- Note: To check multicollinearity, we only need scatterplots, correlations and VIF values between **predictors**.

Variance inflation factor (VIF)

```r
library(car)
vif(m2) # Weight ~ Length + Width
```

```
##   Length    Width
## 20.33948 20.33948
```

```r
vif(m3) # Weight ~ Length + Z
```

```
##   Length        Z
## 531.1215 531.1215
```

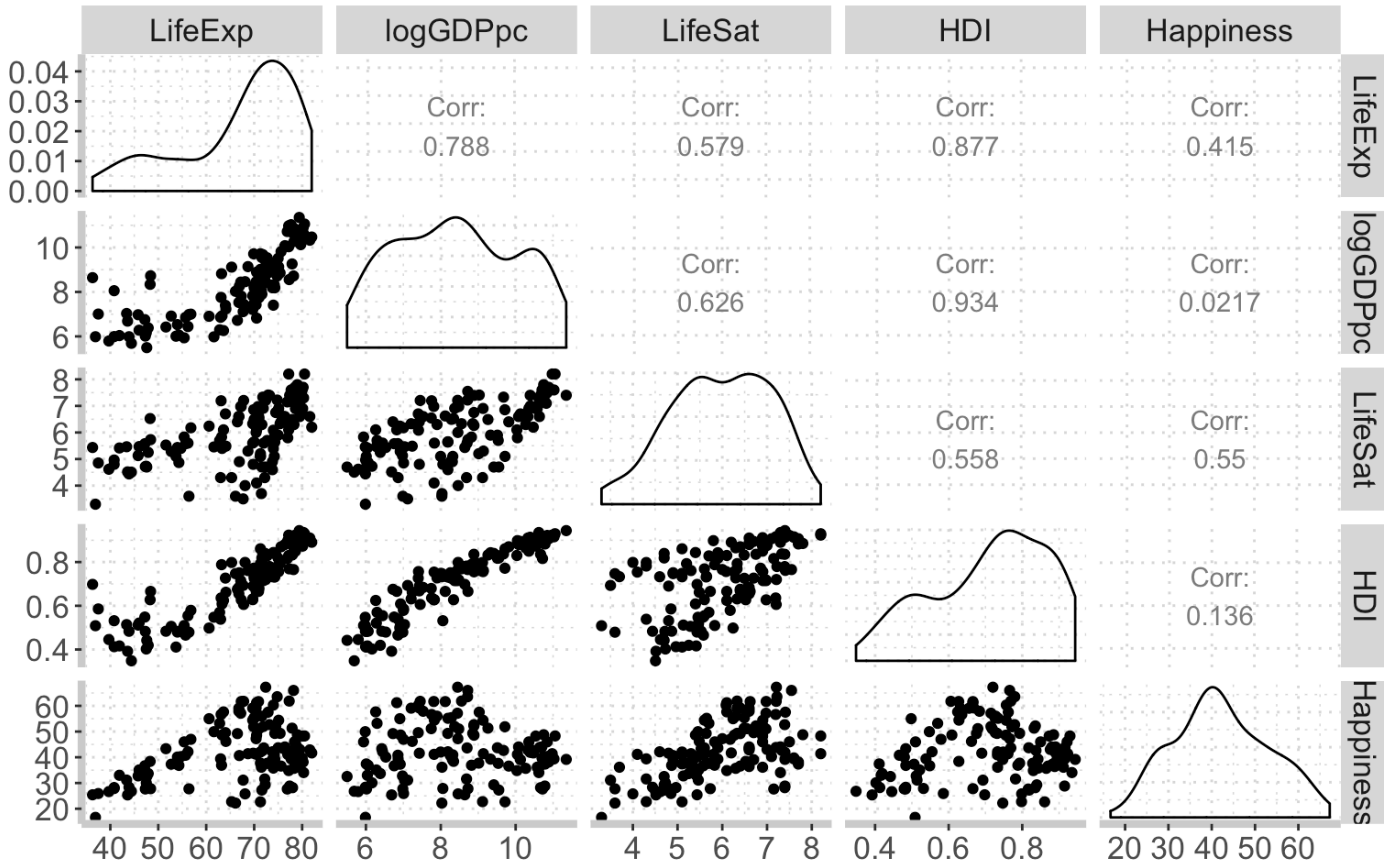
```r
vif(m4) # Weight ~ Length * Width
```

```
##        Length         Width Length:Width
##      25.35773      44.35228     51.43926
```

- All VIF values are much larger than 5. There is multicollinearity existing in these models.

Scatterplot and correlation matrix

```
HappyPlanet <- data.frame(HappyPlanet, logGDPpc=log(HappyPlanet$GDPpc))
ggpairs(data=HappyPlanet[, c("LifeExp","logGDPpc","LifeSat","HDI","Happiness")])
```



- *LifeExp*, *log(GDPpc)* and *HDI* are quite strongly correlated with each other.

Variance inflation factor (VIF)

```r
vif(lm(Happiness ~ LifeExp + logGDPpc + LifeSat + HDI, data=HappyPlanet))
```

```
##   LifeExp  logGDPpc   LifeSat       HDI
##  4.907377  9.537564  1.810730 14.180419
```

```r
vif(lm(Happiness ~ LifeExp * logGDPpc + LifeSat + HDI, data=HappyPlanet))
```

```
##          LifeExp         logGDPpc          LifeSat              HDI
##        44.725475        80.962982         1.819447        15.201240
## LifeExp:logGDPpc
##       171.771083
```

```r
vif(lm(Happiness ~ LifeExp + logGDPpc + LifeSat, data=HappyPlanet))
```

```
##  LifeExp logGDPpc  LifeSat
## 2.731779 2.975119 1.702098
```

```r
vif(lm(Happiness ~ LifeExp * logGDPpc + LifeSat, data=HappyPlanet))
```

```
##          LifeExp         logGDPpc          LifeSat LifeExp:logGDPpc
##        35.362968        59.207237         1.732929       160.236010
```

Some notes

Multicollinearity (high VIF) is **not necessarily a problem**.

You can **ignore i**t when

- The predictors that you are not interested in have high VIF.
- The polynomial terms or interaction terms have high VIF (because these terms are naturally strongly related to the linear terms).
- The dummy variables from the same categorical predictor have high VIF.

Be aware

- Usually, we check multicollinearity in **exploratory data analysis** and only include the **linear terms** of the predictors.
- Multicollinearity causes inflated variance of the estimates, which might lead to insignificant slopes.

Deal with multicollinearity

Solutions to multicollinearity

- Drop one or more predictors.
- Combine some predictors to be one.
- Discount the individual slopes and t tests if you only care about the overall effectiveness of the model.

Model selection criteria

- Nested F test
  - Compare two models with different numbers of predictors to see whether they are significantly different.
- Adjusted R^2
  - Fraction of explained variability (R^2) and the complexity of the model (number of predictors K)
- Mallow's C_p
  - (Un)explained variability, the complexity of the model (number of predictors) and other potential predictors that are not in the model
- Akaike/An information criterion (AIC)
  - Goodness of fit of the model (maximum likelihood value) and the complexity of the model (number of parameters to be estimated)

Mallow's Cp

When evaluating a regression model for a subset of K predictors (current model) from a larger set of m predictors (full model) using a sample of size n, the value of Mallow's C_p is computed by

$$C_p = \frac{SSE_{current}}{MSE_{full}} + 2(K + 1) - n$$

where $SSE_{current}$ is the sum of squared residuals from the current model with K predictors and MSE_{full} is the mean square error for the full model with all m predictors. We prefer models where C_p is small.

- C_p values in different softwares are computed from slightly different formulas. Within a software, choose model with smaller C_p. For example, in R,

$$C_p = SSE_{current} + 2(K + 1)MSE_{full}$$

Mallow's *Cp* in R

```r
HappyPlanet <- HappyPlanet[complete.cases(HappyPlanet), ]
    # make sure that all models have the same sample size
m1 <- lm(Happiness ~ LifeExp+logGDPpc, data=HappyPlanet)
m2 <- lm(Happiness ~ LifeExp+logGDPpc+LifeSat, data=HappyPlanet)
m3 <- lm(Happiness ~ LifeExp+logGDPpc+LifeSat+HDI, data=HappyPlanet)
m4 <- lm(Happiness ~ LifeExp*logGDPpc+LifeSat+HDI, data=HappyPlanet)
anova(m1, m2, m3, m4, test="Cp")
```

```
## Analysis of Variance Table
##
## Model 1: Happiness ~ LifeExp + logGDPpc
## Model 2: Happiness ~ LifeExp + logGDPpc + LifeSat
## Model 3: Happiness ~ LifeExp + logGDPpc + LifeSat + HDI
## Model 4: Happiness ~ LifeExp * logGDPpc + LifeSat + HDI
##   Res.Df    RSS Df Sum of Sq     Cp
## 1    121 8478.9                8597.9
## 2    120 3527.5  1    4951.4 3686.1
## 3    119 3485.6  1      42.0 3683.8
## 4    118 2339.7  1    1145.9 2577.6
```

AIC

Suppose that we have a statistical model with p parameters to be estimated ($p = K + 2$, where K is number of predictors). Let $\hat{L}$ be the maximized value of the likelihood function for the model. Then the AIC value of the model is computed by

$$AIC = 2p - 2ln(\hat{L}).$$

Given a set of candidate models for the data, the preferred model is the one with the minimum AIC value.

```r
AIC(m1, m2, m3, m4)
```

```
##    df      AIC
## m1  4 883.8035
## m2  5 777.0571
## m3  6 777.5733
## m4  7 730.1445
```

Adjusted R-squared, Mallow's *Cp* and AIC

Happiness ~	R^2_{adj}	**Mallow's** C_p	AIC
1. $LifeExp + logGDPpc$	0.426	8597.9	883.8
2. $LifeExp + logGDPpc + LifeSat$	0.759	3686.1	777.1
3. $LifeExp + logGDPpc + LifeSat + HDI$	0.760	3683.8	777.6
4. $LifeExp * logGDPpc + LifeSat + HDI$	0.838	2577.6	730.1

- Model 4 with all four predictors and the interaction between *LifeExp* and *log(GDPpc)* has the highest R^2_{adj} and lowest C_p and AIC.
- Nested F test of model 3 and model 4 suggests model 4 is significantly better.
- If we only compare model 1, 2 and 3, which one is better?
- R^2_{adj} and C_p suggests model 3 while AIC suggests model 2. Since HDI is not significant in model 3, we choose model 2.

The perch models

```
m5 <-  lm(Weight ~ Length + Width, data=perch)
m6 <-  lm(Weight ~ Length + I(Length^2) + Width, data=perch)
m7 <-  lm(Weight ~ Length + Width + I(Width^2), data=perch)
m8 <-  lm(Weight ~ Length + I(Length^2) + Width + I(Width^2), data=perch)
m9 <-  lm(Weight ~ Length * Width, data=perch)
m10 <- lm(Weight ~ Length * Width + I(Length^2), data=perch)
m11 <- lm(Weight ~ Length * Width + I(Width^2), data=perch)
m12 <- lm(Weight ~ Length * Width + I(Length^2) + I(Width^2), data=perch)
AIC(m5,m6,m7,m8,m9,m10,m11,m12)
```

```
##     df      AIC
## m5   4 666.1566
## m6   5 589.2590
## m7   5 599.3550
## m8   6 588.2196
## m9   5 589.2048
## m10  6 589.8127
## m11  6 591.2017
## m12  7 588.1613
```

The perch models

```r
anova(m5,m6,m7,m8,m9,m10,m11,m12, test="Cp")
```

```
## Analysis of Variance Table
##
## Model 1: Weight ~ Length + Width
## Model 2: Weight ~ Length + I(Length^2) + Width
## Model 3: Weight ~ Length + Width + I(Width^2)
## Model 4: Weight ~ Length + I(Length^2) + Width + I(Width^2)
## Model 5: Weight ~ Length * Width
## Model 6: Weight ~ Length * Width + I(Length^2)
## Model 7: Weight ~ Length * Width + I(Width^2)
## Model 8: Weight ~ Length * Width + I(Length^2) + I(Width^2)
##   Res.Df    RSS Df Sum of Sq     Cp
## 1     53 416762                427922
## 2     52 101863  1    314899 116743
## 3     52 121987  0    -20124 136867
## 4     51  96482  1     25505 115082
## 5     52 101765 -1     -5283 116645
## 6     51  99266  1      2499 117866
## 7     51 101759  0     -2493 120359
## 8     50  93000  1      8759 115320
```

The perch models

Weight ~	R^2_{adj}	C_p	AIC
5. $L + W$	0.9349	427,922	666.16
6. $L + L^2 + W$	0.9838	116,743	589.26
7. $L + W + W^2$	0.9806	136,867	599.36
8. $L + L^2 + W + W^2$	0.9843	**115,082**	588.22
9. $L + W + LW$	0.9838	116,645	589.20
10. $L + L^2 + W + LW$	0.9839	117,866	589.81
11. $L + W + W^2 + LW$	0.9835	120,359	591.20
12. $L + L^2 + W + W^2 + LW$	**0.9846**	115,320	**588.16**

- R^2_{adj} and C_p suggest model 12. AIC suggests model 8. Nested F test suggests model 6 or 9.
- The choice depends on the purpose of modeling. If we want to explain as much variability as possible, model 12; if we want better model fitting in terms of likelihood value, model 8; if we want a model that is easier to interpret, model 6 or 9.

Summary

Multicollinearity

- Definition and examples
- Detect multicollinearity
  - Scatterplot matrix
  - Correlation matrix
  - Variance inflation factor (VIF)
- Deal with multicollinearity

Model selection criteria

- Nested F test
- Adjusted R^2
- Mallow's C_p
- Akaike/An information criterion (AIC)