STAT021 Statistical Methods II

Lecture 18 Transforming Predictors

Lu Chen
Swarthmore College
11/13/2018

Outline

The purpose of transforming the predictors is to achieve better model fitting, i.e., to explain as much variability in the reponse variable as possible while at the same time to keep the model simple.

There are many different ways to transform the predictors, such as

- Applying a function on the predictor (e.g., natrual logarithm function)
- **Categorizing a quantitative predictor**
- **Treating a categorical predictor quantitative**
- **Including the polynomial terms of a quantitative predictor**
- ...

Categorization is common

```r
head(HappyPlanet)
```

```
##               Country Happiness      GDPpc       HDI HDI2     HDI4
## 1         Philippines  59.17430  1678.8520 0.6682183  Low   Medium
## 2              Rwanda  28.34747   398.2085 0.4832405  Low      Low
## 3             Hungary  37.63759 13842.6055 0.8283505 High VeryHigh
## 4              Cyprus  45.99012 31386.6326 0.8497454 High VeryHigh
## 5 Trinidad and Tobago  51.86844 16530.1804 0.7718938 High     High
## 6            Paraguay  51.12862  2312.1925 0.6791644  Low   Medium
```

To categorize a quantitative variable, the cutoffs may be chosen by

- quantiles
- practical meaning
- model fitting (try different cutoffs and choose the set that results in best model fitting)
- convenience

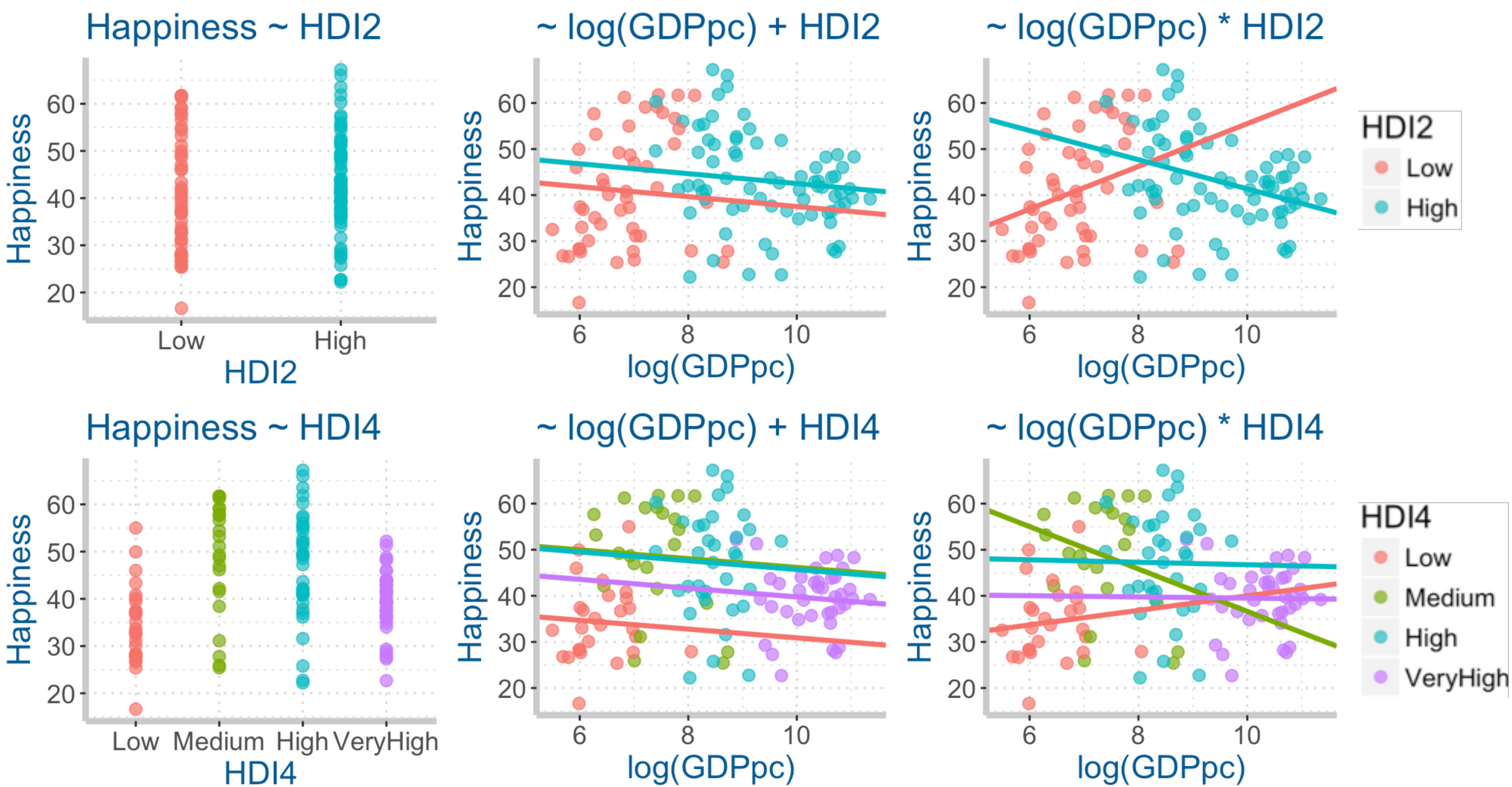
Categorization is common

Example: the Breast Cancer Risk Assessment Tool for breast cancer prediction has the following predictors

- Age (< 50, ≥ 50)
- Age at first period (7~11, 12~13, ≥ 14 or unknown)
- Age at first live birth (< 20 or unknown, 20~24, 25~29 or never, ≥ 30)
- Number of first-degree relatives with breast cancer (0 or unknown, 1, > 1)
- Number of past breast biopsies (0 or unknown, 1 or unknown but with positive, > 1)
- Race/ethnicity
- Breast density (0~0.25, 0.25~0.5, 0.5~0.75, 0.75~1)

Why categorization?

Comparing the six models - best model?



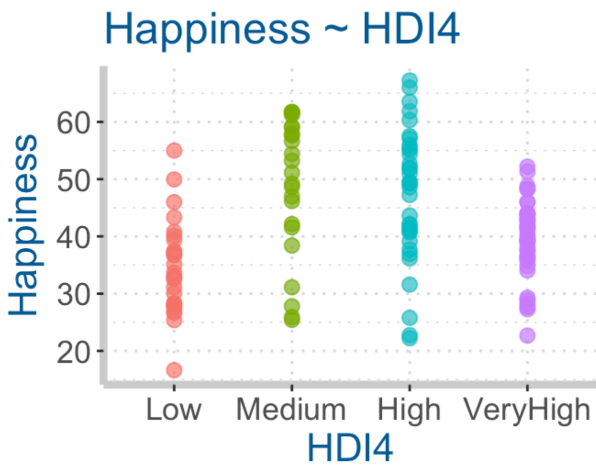
Comparing the six models

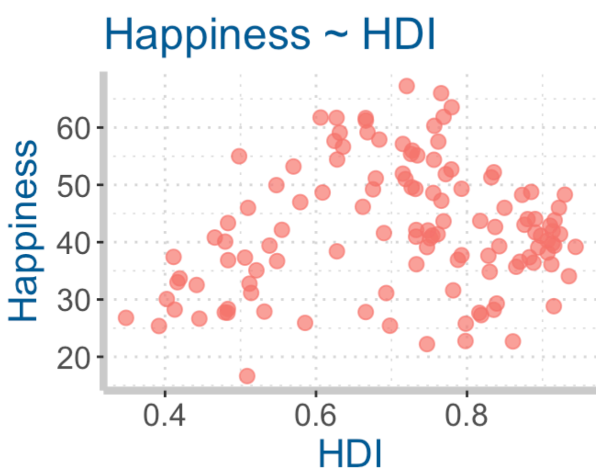
Happiness ~	F **test**	R^2	R^2_{adj}
1. *HDI2*	$F = 1.2, P = 0.281$	0.0095	0.0014
2. *log(GDPpc) + HDI2*	$F = 1.1, P = 0.328$	0.0183	0.0021
3. *log(GDPpc) * HDI2*	$F = 4.9, P = 0.003$	0.1088	0.0865
4. *HDI4*	$F = 13.5, P = 1.21 \times 10^{-7}$	0.2522	0.2335
5. *log(GDPpc) + HDI4*	$F = 10.2, P = 4.13 \times 10^{-7}$	0.2546	0.2295
6. *log(GDPpc) * HDI4*	$F = 6.1, P = 4.82 \times 10^{-7}$	0.2680	0.2238

Note: In model 5, *log(GDPpc)* is not significant; In model 6, *log(GDPpc)* and the interaction terms are not significant.

- Model 4 (*Happiness* ~ *HDI4*) is the best one based on adjusted R^2 and nested F tests.

Why categorizing predictors?

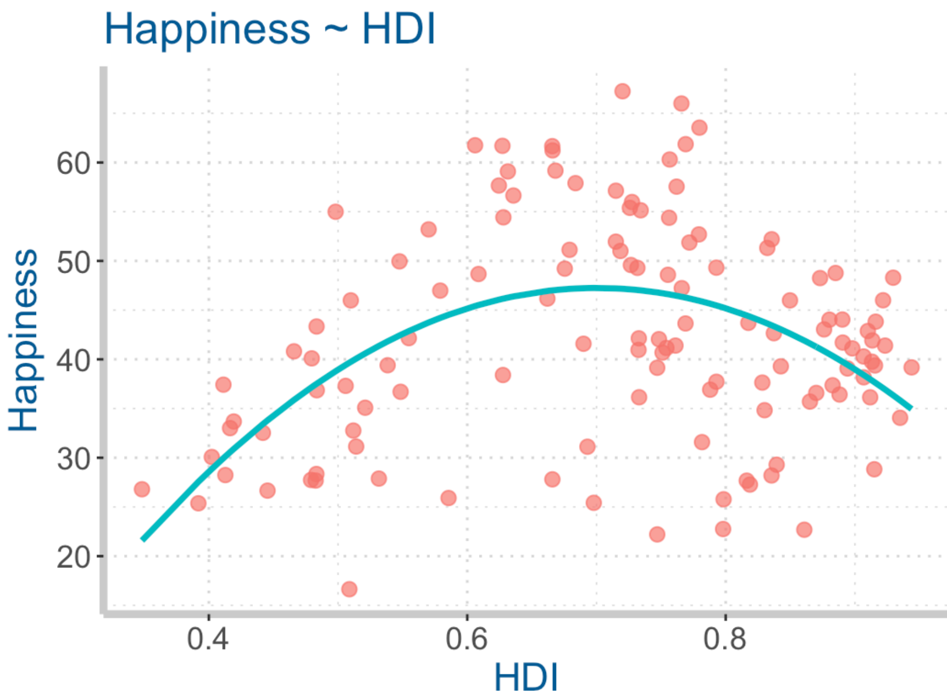




Happiness ~	F test	R^2	R^2_{adj}
HDI4	$F = 13.5, P = 1.21 \times 10^{-7}$	0.2522	0.2335
HDI	$F = 2.68, P = 0.10$	0.0215	0.0135

- Linear regression of *Happiness* and *HDI* assumes linear relationship, which is in fact a very strong assumption.
- However, coding *HDI* as a categorical variable with four categories assumes means of *Happiness* are different for the four *HDI* levels.
- Categorizing a predictor allows more flexibility in model assumptions about the relationship between the response variable and the predictor but adds more complexity (more parameters) to the model.

Another way to model *Happiness* ~ *HDI*



- The relationship between *Happiness* and *HDI*: as *HDI* increases, *Happiness* first goes up and then goes down - non-linear.
- Polynomial Regression

Polynomial regression

56 perches were caught in a lake in Finland and three variables were measured.

- *Weight* (in grams)
- *Length* (in centimeters)
- *Width* (in centimeters)

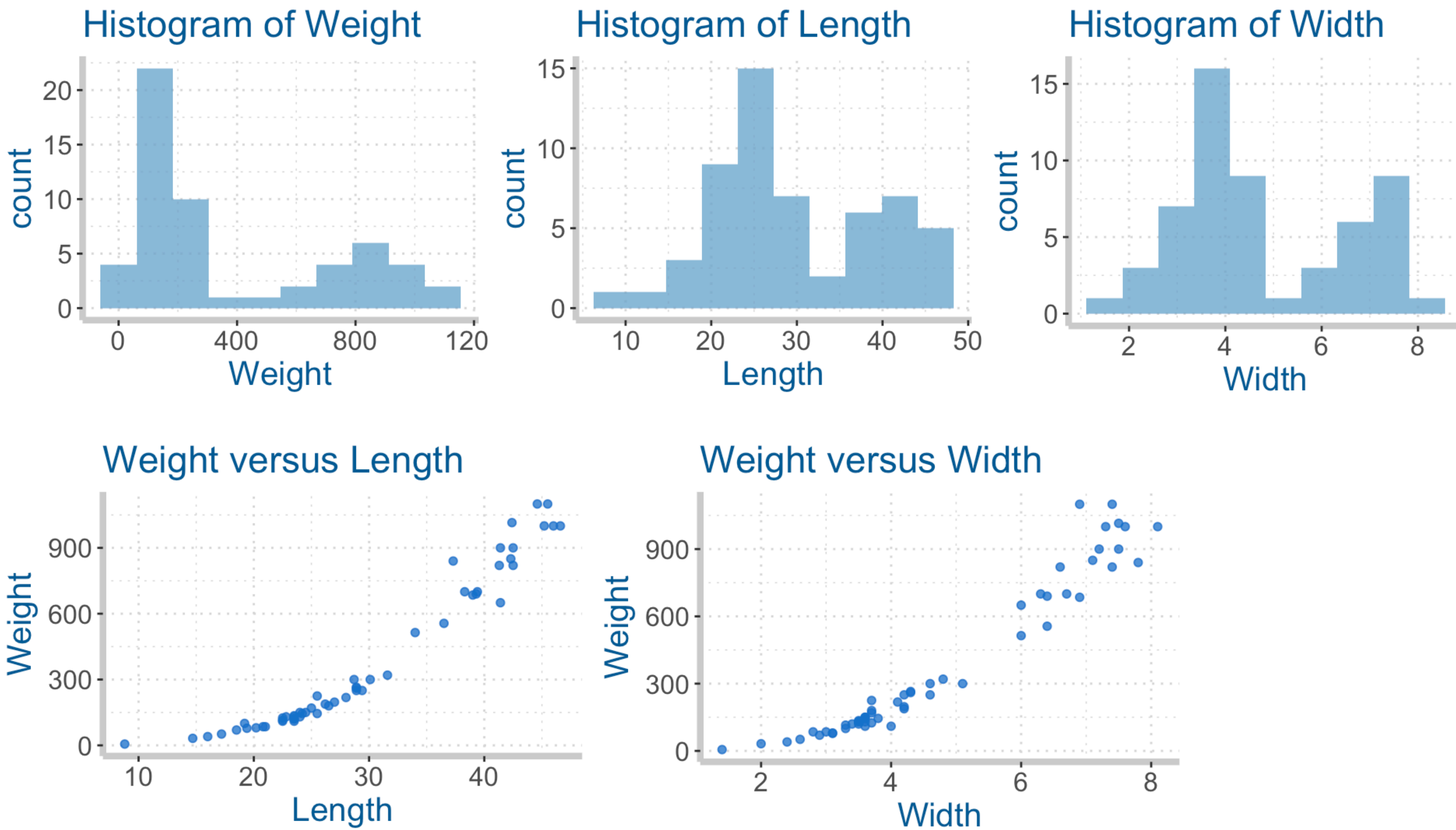
```
head(perch, 3)
```

```
##   Weight Length Width
## 1    5.9    8.8   1.4
## 2   32.0   14.7   2.0
## 3   40.0   16.0   2.4
```

```
library(psych) # package for the describe() function
describe(perch)[,2:4] # summary statistics
```

```
##         n   mean     sd
## Weight 56 382.24 347.62
## Length 56  29.57   9.53
## Width  56   4.74   1.78
```

Polynomial regression



Polynomial regression

For a single quantitative predictor X, a **polynomial regression model** of degree k has the form

$$Y = \beta_0 + \beta_1 X + \beta_2 X^2 + \cdots + \beta_k X^k + \epsilon, \epsilon \overset{iid}{\sim} N(0, \sigma)$$

For a single quantitative predictor X, a **quadratic regression model** has the form

$$Y = \beta_0 + \beta_1 X + \beta_2 X^2 + \epsilon, \epsilon \overset{iid}{\sim} N(0, \sigma)$$

- We do not consider X and X^2 as two different predictors. Instead, we assume the relationship between Y and the single predictor X is nonlinear and quadratic.
- Usually, if the t test for the slope of the quadratic term X^2 is NOT significant, we may remove the quadratic term from the model (or try a higher order term).
- However, as long as the highest order term is significant, we keep all the lower order terms.

Polynomial regression

Response variable: *Weight*
Predictors: *Length* (L), *Width* (W)

We will consider 12 different models that involve *Length* and *Width*.

> **Complete second-order model** is the model that includes linear and quadratic terms for both predictors along with the interaction term.

1. $Weight \sim L$
2. $Weight \sim L + L^2$
3. $Weight \sim W$
4. $Weight \sim W + W^2$
5. $Weight \sim L + W$
6. $Weight \sim L + L^2 + W$
7. $Weight \sim L + W + W^2$
8. $Weight \sim L + L^2 + W + W^2$
9. $Weight \sim L + W + LW$
10. $Weight \sim L + L^2 + W + LW$
11. $Weight \sim L + W + W^2 + LW$
12. $Weight \sim L + L^2 + W + W^2 + LW$

Polynomial regression

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

```
##             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 ***
##
## 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
```

```r
summary(m2 <- lm(Weight ~ Length + I(Length^2), data=perch))
```

```
##              Estimate Std. Error t value Pr(>|t|)
## (Intercept) 128.34533   78.77870   1.629  0.10920
## Length      -21.02388    5.41770  -3.881  0.00029 ***
## I(Length^2)   0.90862    0.08689  10.458 1.72e-14 ***
##
## Residual standard error: 56.99 on 53 degrees of freedom
## Multiple R-squared:  0.9741, Adjusted R-squared:  0.9731
## F-statistic: 996.6 on 2 and 53 DF,  p-value: < 2.2e-16
```

Polynomial regression

```r
summary(lm(Weight ~ Length + I(Length^2) + I(Length^3), data=perch))
```

```
##                Estimate Std. Error t value Pr(>|t|)
## (Intercept) 304.981990 160.823863   1.896   0.0635 .
## Length      -43.397306  18.587729  -2.335   0.0235 *
## I(Length^2)   1.769911   0.690263   2.564   0.0133 *
## I(Length^3)  -0.010149   0.008069  -1.258   0.2141
##
## Residual standard error: 56.68 on 52 degrees of freedom
## Multiple R-squared:  0.9749, Adjusted R-squared:  0.9734
## F-statistic: 672.2 on 3 and 52 DF,  p-value: < 2.2e-16
```

- The cubic term is not significant and thus not necessary in the model.
- The model

$$\widehat{Weight} = 128.3 - 21.0 \times Length + 0.9 \times Length^2$$

seems the best among the three.

Polynomial regression

```r
summary(m3 <- lm(Weight ~ Width, data=perch))
```

```
##             Estimate Std. Error t value Pr(>|t|)
## (Intercept) -509.289     35.594  -14.31   <2e-16 ***
## Width        188.115      7.038   26.73   <2e-16 ***
##
## Residual standard error: 93 on 54 degrees of freedom
## Multiple R-squared:  0.9297, Adjusted R-squared:  0.9284
## F-statistic: 714.5 on 1 and 54 DF,  p-value: < 2.2e-16
```

```r
summary(m4 <- lm(Weight ~ Width + I(Width^2), data=perch))
```

```
##             Estimate Std. Error t value Pr(>|t|)
## (Intercept)  -31.735     91.908  -0.345    0.731
## Width        -25.635     39.487  -0.649    0.519
## I(Width^2)    20.934      3.827   5.470 1.24e-06 ***
##
## Residual standard error: 75.05 on 53 degrees of freedom
## Multiple R-squared:  0.9551, Adjusted R-squared:  0.9534
## F-statistic: 563.5 on 2 and 53 DF,  p-value: < 2.2e-16
```

- Because the quadratic term is significant, although the linear term is not, we still keep both terms in the model.

Polynomial regression

```r
summary(m5 <- lm(Weight ~ Length + Width, data=perch))
```

```
##             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 ***
##
## Residual standard error: 88.68 on 53 degrees of freedom
## Multiple R-squared:  0.9373, Adjusted R-squared:  0.9349
## F-statistic: 396.1 on 2 and 53 DF,  p-value: < 2.2e-16
```

- *Length* and *Width* each has a significant quadratic relationship with *Weight*. Let's now evaluate models with both predictors, starting from the linear terms.
- Given *Width* (*Length*) is held constant, *Weight* and *Length* (*Width*) has a significant linear relationship.
- But this model explains slightly less variability ($R^2 = 0.94$) than the previous quadratic models ($R^2 = 0.97$ and 0.96).

Polynomial regression

```
summary(m6 <- lm(Weight ~ Length + I(Length^2) + Width, data=perch))
```

```
##               Estimate Std. Error t value Pr(>|t|)
## (Intercept) 147.12090   61.25958   2.402   0.0199 *
## Length      -34.71805    4.78840  -7.250 1.97e-09 ***
## I(Length^2)   0.86134    0.06794  12.679  < 2e-16 ***
## Width        91.09772   15.20858   5.990 2.00e-07 ***
##
## Residual standard error: 44.26 on 52 degrees of freedom
## Multiple R-squared:  0.9847, Adjusted R-squared:  0.9838
## F-statistic:  1114 on 3 and 52 DF,  p-value: < 2.2e-16
```

Polynomial regression

```r
summary(m7 <- lm(Weight ~ Length + Width + I(Width^2), data=perch))
```

```
##              Estimate Std. Error t value Pr(>|t|)
## (Intercept)    36.348     59.832   0.607    0.546
## Length         29.176      3.363   8.674 1.11e-11 ***
## Width        -271.668     38.130  -7.125 3.13e-09 ***
## I(Width^2)     30.128      2.688  11.210 1.74e-15 ***
##
## Residual standard error: 48.43 on 52 degrees of freedom
## Multiple R-squared:  0.9816, Adjusted R-squared:  0.9806
## F-statistic:   927 on 3 and 52 DF,  p-value: < 2.2e-16
```

Polynomial regression

```r
summary(m8 <- lm(Weight ~ Length+I(Length^2)+Width+I(Width^2), data=perch))
```

```
##              Estimate Std. Error t value Pr(>|t|)
## (Intercept) 138.0015    60.4435   2.283 0.026621 *
## Length      -15.2436    12.4688  -1.223 0.227124
## I(Length^2)   0.6065     0.1652   3.672 0.000577 ***
## Width       -31.0365    73.9416  -0.420 0.676436
## I(Width^2)   10.0718     5.9717   1.687 0.097793 .
##
## Residual standard error: 43.49 on 51 degrees of freedom
## Multiple R-squared:  0.9855, Adjusted R-squared:  0.9843
## F-statistic: 865.5 on 4 and 51 DF,  p-value: < 2.2e-16
```

- From model 6 to 8, the added term $Width^2$ term is not/marginally significant. R^2 increases only a little bit. It is probably not necessary to include $Width^2$.
- From model 7 to 8, the added term $Length^2$ is significant. We probably will keep it.

Model comparisons

Weight ~	F	R^2	R^2_{adj}
1. L	626.5	0.9207	0.9192
2. $L + L^2$	996.6	0.9741	0.9731
3. W	714.5	0.9297	0.9284
4. $W + W^2$	563.5	0.9551	0.9534
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

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

- Model 8 has the largest R^2 and adjusted R^2.
- Comparing model 6 and 8, the $Width^2$ is not significant suggesting that model 8 is not significantly better than model 6.
- Therefore, we choose model 6 as the best model among the 8 models.
- Note model 7 also has comparable assessment as model 6.

Model comparisons

```r
summary(m8)$coefficents
```

```
##               Estimate Std. Error t value Pr(>|t|)
## (Intercept) 138.0015    60.4435   2.283 0.026621 *
## Length      -15.2436    12.4688  -1.223 0.227124
## I(Length^2)   0.6065     0.1652   3.672 0.000577 ***
## Width       -31.0365    73.9416  -0.420 0.676436
## I(Width^2)   10.0718     5.9717   1.687 0.097793 .
```

```r
anova(m6, m8)
```

```
## Analysis of Variance Table
##
## Model 1: Weight ~ Length + I(Length^2) + Width
## Model 2: Weight ~ Length + I(Length^2) + Width + I(Width^2)
##   Res.Df    RSS Df Sum of Sq      F  Pr(>F)
## 1     52 101863
## 2     51  96482  1    5381.3 2.8445 0.09779 .
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

Model comparisons

```
summary(m9 <- lm(Weight ~ Length * Width, data=perch))
```

```
##                Estimate Std. Error t value Pr(>|t|)
## (Intercept)    113.9349    58.7844   1.938    0.058 .
## Length          -3.4827     3.1521  -1.105    0.274
## Width          -94.6309    22.2954  -4.244 9.06e-05 ***
## Length:Width     5.2412     0.4131  12.687  < 2e-16 ***
##
## Residual standard error: 44.24 on 52 degrees of freedom
## Multiple R-squared:  0.9847, Adjusted R-squared:  0.9838
## F-statistic:  1115 on 3 and 52 DF,  p-value: < 2.2e-16
```

- The interaction term $Length \times Width$ is highly significant. Keep it.

Model comparisons

```
summary(m10 <- lm(Weight ~ Length * Width + I(Length^2), data=perch))
```

```
##                Estimate Std. Error t value Pr(>|t|)
## (Intercept)    136.1045    61.8038   2.202   0.0322 *
## Length         -19.2196    14.2406  -1.350   0.1831
## Width           -3.5967    83.3663  -0.043   0.9658
## I(Length^2)      0.4300     0.3795   1.133   0.2625
## Length:Width     2.6672     2.3089   1.155   0.2534
##
## Residual standard error: 44.12 on 51 degrees of freedom
## Multiple R-squared:  0.9851, Adjusted R-squared:  0.9839
## F-statistic: 840.9 on 4 and 51 DF,  p-value: < 2.2e-16
```

- The added quadratic term $Length^2$ is not significant when the interaction term is already in the model. We may not need it.

Model comparisons

```r
summary(m11 <- lm(Weight ~ Length * Width + I(Width^2), data=perch))
```

```
##                Estimate Std. Error t value Pr(>|t|)
## (Intercept)    114.5331    60.3972   1.896  0.06359 .
## Length          -4.0403    10.8834  -0.371  0.71200
## Width          -91.2681    66.6836  -1.369  0.17710
## I(Width^2)      -0.5327     9.9434  -0.054  0.95748
## Length:Width     5.3281     1.6734   3.184  0.00248 **
##
## Residual standard error: 44.67 on 51 degrees of freedom
## Multiple R-squared:  0.9847, Adjusted R-squared:  0.9835
## F-statistic:   820 on 4 and 51 DF,  p-value: < 2.2e-16
```

- The added quadratic term $Width^2$ is not significant, either, when the interaction term is already in the model. We may not need it.

Model comparisons

```r
summary(m12 <- lm(Weight ~ Length*Width+I(Length^2)+I(Width^2), data=perch))
```

```
##                 Estimate Std. Error t value Pr(>|t|)
## (Intercept)     156.3486    61.4152   2.546   0.0140 *
## Length          -25.0007    14.2729  -1.752   0.0860 .
## Width            20.9772    82.5877   0.254   0.8005
## I(Length^2)       1.5719     0.7244   2.170   0.0348 *
## I(Width^2)       34.4058    18.7455   1.835   0.0724 .
## Length:Width     -9.7763     7.1455  -1.368   0.1774
##
## Residual standard error: 43.13 on 50 degrees of freedom
## Multiple R-squared:  0.986,  Adjusted R-squared:  0.9846
## F-statistic: 704.6 on 5 and 50 DF,  p-value: < 2.2e-16
```

- With both quadratic terms ($Length^2$ and $Width^2$) added, the interaction term becomes insignificant.
- The inferences of model 10, 11 and 12 suggest that we may only need one of the three terms: $Length^2$, $Width^2$ and $Length \times Width$.

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. Is it the best model?
- Model 10 and 11 are not significantly better than model 9.
- Model 9 has similar R^2 and adjusted R^2 as model 6 and model 12. Let's compare them using nested F test.

Model comparisons

```r
anova(m6, m12)
```

```
## Analysis of Variance Table
##
## Model 1: Weight ~ Length + I(Length^2) + Width
## Model 2: Weight ~ Length * Width + I(Length^2) + I(Width^2)
##   Res.Df    RSS Df Sum of Sq      F Pr(>F)
## 1     52 101863
## 2     50  93000  2    8863.1 2.3825 0.1027
```

- Model 12 is NOT significantly better than model 6.

```r
anova(m9, m12)
```

```
## Analysis of Variance Table
##
## Model 1: Weight ~ Length * Width
## Model 2: Weight ~ Length * Width + I(Length^2) + I(Width^2)
##   Res.Df    RSS Df Sum of Sq      F Pr(>F)
## 1     52 101765
## 2     50  93000  2    8764.6 2.3561 0.1052
```

- Model 12 is NOT significantly better than model 9.

Model comparisons

```r
anova(m6, m9)
```

```
## Analysis of Variance Table
##
## Model 1: Weight ~ Length + I(Length^2) + Width
## Model 2: Weight ~ Length * Width
##   Res.Df    RSS Df Sum of Sq F Pr(>F)
## 1      52 101863
## 2      52 101765  0    98.527
```

- Nested F test cannot be used to compare two models with the same number of predictors.
- Model 6 and model 9 have very close R^2 and adjusted R^2. Both models are good.
- This output gives the `RSS` (residual sum of squares, SSE) values of the two models. A model with smaller SSE explains more variablity in the response variable than the model with larger SSE.
- Therefore, model 9 is slightly-slightly better than model 6.

Happiness ~ HDI

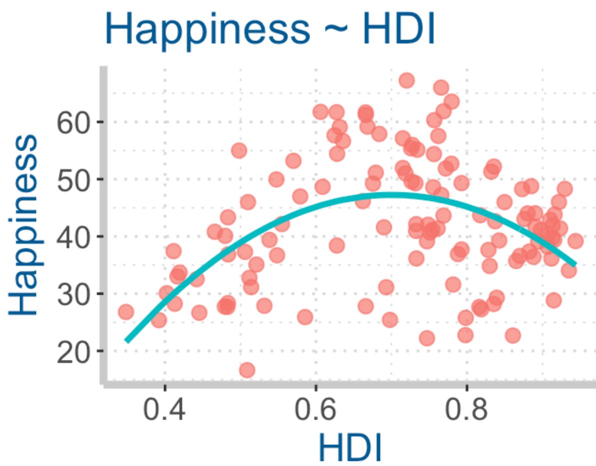
```r
summary(m.quan <- lm(Happiness ~ HDI, data=HappyPlanet))
```

```
##             Estimate Std. Error t value Pr(>|t|)
## (Intercept)   34.834      4.617   7.545 9.01e-12 ***
## HDI           10.379      6.340   1.637    0.104
##
## Residual standard error: 10.98 on 122 degrees of freedom
## Multiple R-squared:  0.02149,    Adjusted R-squared:  0.01347
## F-statistic: 2.679 on 1 and 122 DF,  p-value: 0.1042
```

```r
summary(m.cate <- lm(Happiness ~ HDI4, data=HappyPlanet))
```

```
##              Estimate Std. Error t value Pr(>|t|)
## (Intercept)    34.315      1.862  18.429  < 2e-16 ***
## HDI4Medium     14.337      2.745   5.222 7.53e-07 ***
## HDI4High       12.838      2.478   5.180 9.05e-07 ***
## HDI4VeryHigh    5.146      2.422   2.124   0.0357 *
##
## Residual standard error: 9.675 on 120 degrees of freedom
## Multiple R-squared:  0.2522, Adjusted R-squared:  0.2335
## F-statistic: 13.49 on 3 and 120 DF,  p-value: 1.213e-07
```

Happiness ~ HDI



$$\widehat{Happiness} = -54.3 + 290.2 \times HDI - 207.2 \times HDI^2$$

Happiness ~	F **test**	R^2	R^2_{adj}
HDI	$F = 2.68, P = 0.10$	0.0215	0.0135
HDI4	$F = 13.5, P = 1.21 \times 10^{-7}$	0.2522	0.2335
HDI+HDI^2	$F = 16.6, P = 4.37 \times 10^{-7}$	0.215	0.202

```
summary(m.poly <- lm(Happiness ~ HDI + I(HDI^2), data=HappyPlanet))
```

```
##              Estimate Std. Error t value Pr(>|t|)
## (Intercept)   -54.34      16.85  -3.225  0.00162 **
## HDI           290.16      51.55   5.629 1.20e-07 ***
## I(HDI^2)     -207.19      37.94  -5.461 2.57e-07 ***
##
## Residual standard error: 9.872 on 121 degrees of freedom
## Multiple R-squared:  0.215,  Adjusted R-squared:  0.202
## F-statistic: 16.57 on 2 and 121 DF,  p-value: 4.372e-07
```

Summary

- Transforming the predictors:
  - Applying a function on the predictor
  - **Categorizing a quantitative predictor**
  - **Treating a categorical predictor quantitative**
  - **Including the polynomial terms of a quantitative predictor**
  - ...
- Nested F test and adjusted R^2 are used together to select best model.
- Note: the model selecting process in this lecture did not consider model assumptions. In a full analysis, one should always check whether a model's assumptions are satisfied when selecting models.