228.371 - Statistical Modelling for Engineers and Technologists

Week 3. Polynomial and Multiple Regression

Dr. Daniel Walsh
IIMS 3.07 x 41032
d.c.walsh@massey.ac.nz

Semester One - 2015

Polynomials

- Additional powers of x can be included in a linear model:

$$y_i = \beta_0 + \beta_1 x_i + \beta_2 x_i^2 + \beta_3 x_i^3 + \ldots + \epsilon_i$$

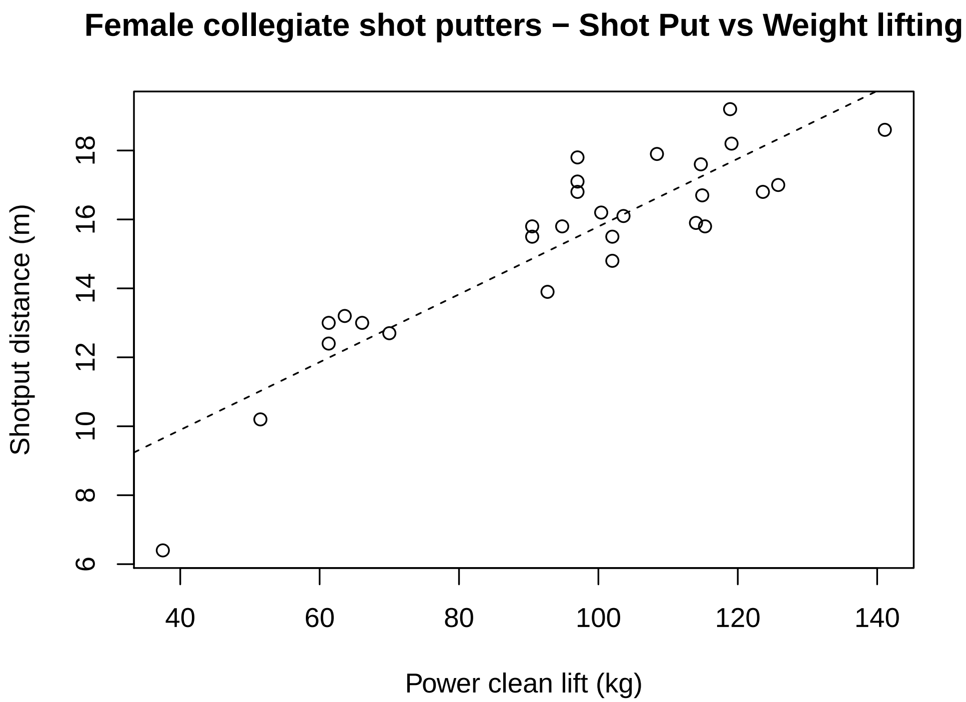
- Assumptions of normality and constant variance still need to be satisfied by y.
- The powers of x are entered in order and are tested for significance in that order.
- In practical terms additional variables equal to x^2, x^3 etc are calculated and added to the model.

Example: Shotput.csv

```
Shot <- read.csv (file="Data/Shotput.csv", header=TRUE)
```

power.clean	shot.putt
37.5	6.4
51.5	10.2
61.3	12.4
61.3	13
63.6	13.2
66.1	13
70	12.7
92.7	13.9
90.5	15.5
90.5	15.8
94.8	15.8
97	16.8
97	17.1
97	17.8
102	14.8
102	15.5
103.6	16.1
100.4	16.2
108.4	17.9
114	15.9
115.3	15.8
114.9	16.7
114.7	17.6
123.6	16.8
125.8	17
119.1	18.2
118.9	19.2
141.1	18.6

Example: Shotput.csv

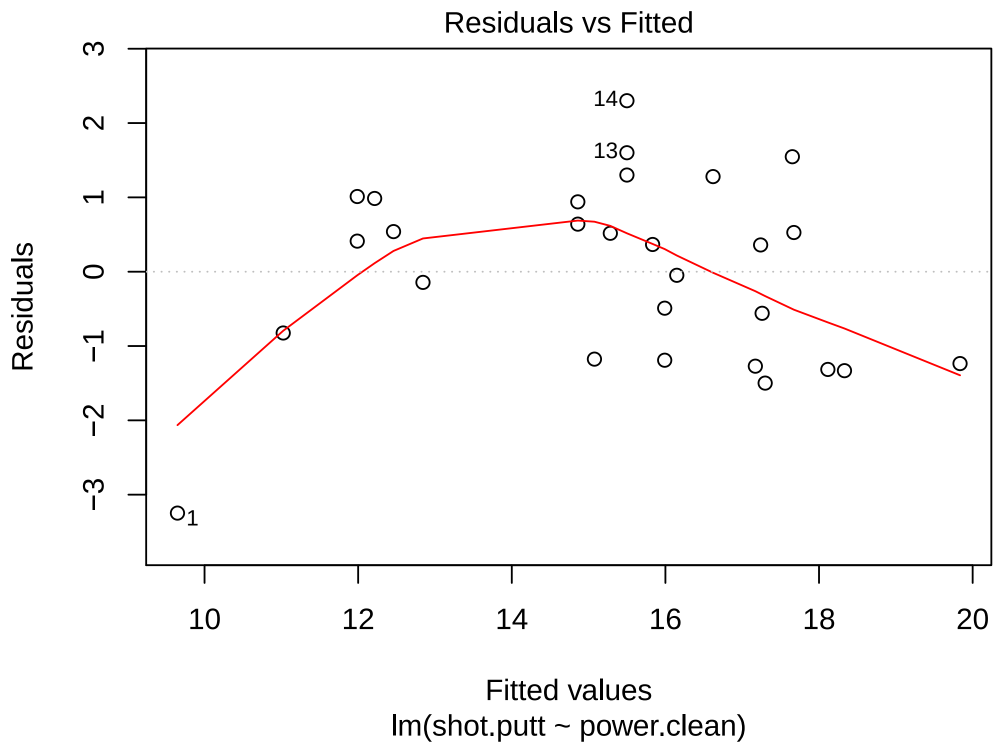



```
m1 <- lm (shot.putt ~ power.clean, data=Shot)
summary(m1)
```

```
Call:
lm(formula = shot.putt ~ power.clean, data = Shot)

Residuals:
    Min      1Q  Median      3Q     Max
-3.2475 -1.1798  0.3635  0.9516  2.3010

Coefficients:
            Estimate Std. Error t value Pr(>|t|)
(Intercept) 5.959629   0.958835   6.215 1.42e-06 ***
power.clean 0.098344   0.009721  10.117 1.66e-10 ***
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 1.265 on 26 degrees of freedom
Multiple R-squared:  0.7974, Adjusted R-squared:  0.7896
F-statistic: 102.4 on 1 and 26 DF,  p-value: 1.663e-10
```

Example: `Shotput.csv` - Residual Plot

```
plot(m1, 1)
```



- Residual plot show some evidence of curvature.
- Try a quadratic model.

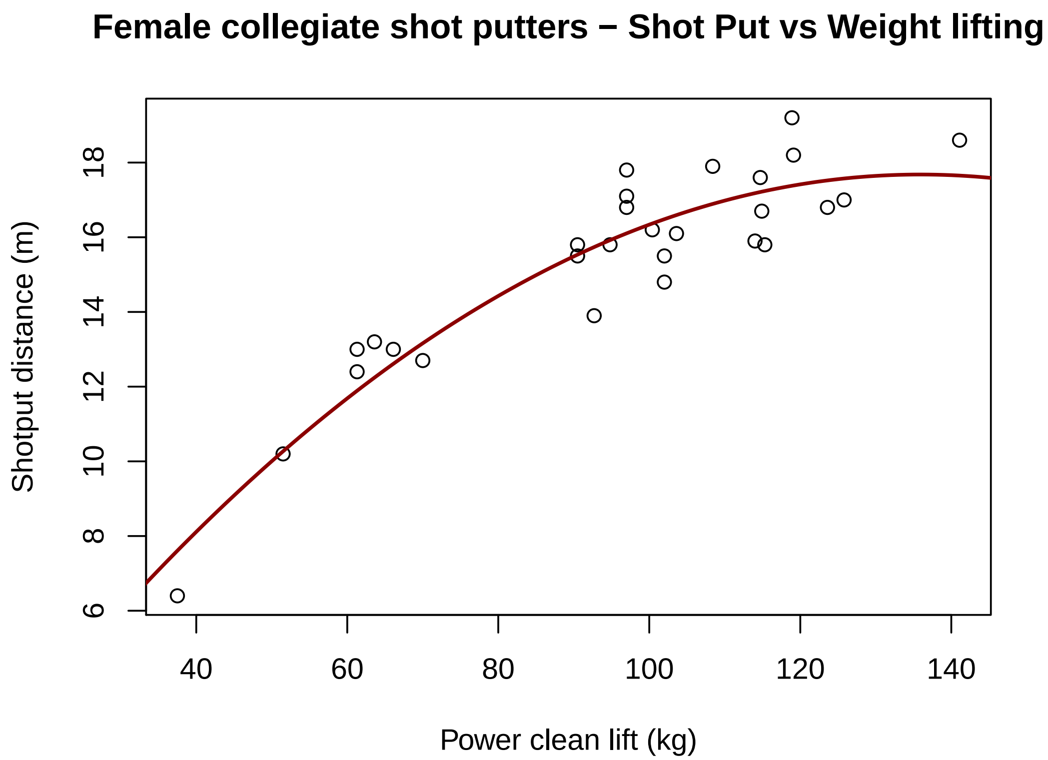
Example: Shotput.csv - Quadratic Model

```
m2 <- lm (shot.putt ~ power.clean + I(power.clean^2), data=Shot)
summary(m2)
```

```
Call:
lm(formula = shot.putt ~ power.clean + I(power.clean^2), data = Shot)

Residuals:
     Min       1Q   Median       3Q      Max 
-1.83778 -0.61059 -0.05209  0.84770  1.82122 

Coefficients:
                   Estimate Std. Error t value Pr(>|t|)    
(Intercept)      -1.5318210  2.1813993  -0.702  0.48903    
power.clean       0.2827080  0.0506977   5.576 8.46e-06 ***
I(power.clean^2) -0.0010400  0.0002824  -3.682  0.00111 ** 
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 1.039 on 25 degrees of freedom
Multiple R-squared:  0.8687, Adjusted R-squared:  0.8582 
F-statistic: 82.68 on 2 and 25 DF,  p-value: 9.54e-12
```

- The coefficient of power.clean^2 is significant.

Example: Shotput.csv - Quadratic Model




```
plot(Shot$shot.putt ~ Shot$power.clean,
     xlab="Power clean lift (kg)",
     ylab="Shotput distance (m)",
     main="Female collegiate shot putters - Shot Put vs Weight lifting")
x <- seq (30,150)
y <- predict.lm (m2, data.frame(power.clean=x))
points (y ~ x, type="l", lwd=2, col="red4")
```

Comparing polynomial models

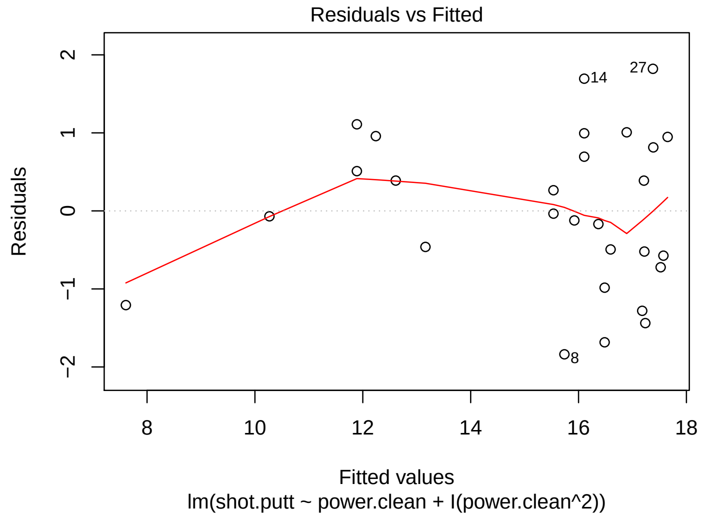
- Each model needs to be fitted in turn.
- You will always include every term of lower order in each model. In a cubic you must include quadratic and linear terms.
- Compare using residual error, adjusted R^2 (more later), and the standard residual analysis graphics.

Linear model: `Shotput.csv`

```
Residual standard error: 1.265 on 26 degrees of freedom
Multiple R-squared:  0.7974, Adjusted R-squared:  0.7896
F-statistic: 102.4 on 1 and 26 DF,  p-value: 1.663e-10
```

Quadratic model: `Shotput.csv`

```
Residual standard error: 1.039 on 25 degrees of freedom
Multiple R-squared:  0.8687, Adjusted R-squared:  0.8582
F-statistic: 82.68 on 2 and 25 DF,  p-value: 9.54e-12
```




```
plot(m2, 1)
```

Comparing polynomail models - F-test

- The linear model is a subset of the quadratic model (if the coefficient of the quadratic model is zero it is a linear model).
- Because of this we can compare the models with an F-test.
- H_0 : "Linear (smaller) model is correct" vs H_a : "Quadratic model is correct".

```
anova (m1, m2)
```

```
Analysis of Variance Table

Model 1: shot.putt ~ power.clean
Model 2: shot.putt ~ power.clean + I(power.clean^2)
  Res.Df    RSS Df Sum of Sq      F   Pr(>F)
1     26 41.637
2     25 26.995  1    14.643 13.561 0.001114 **
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

- The p-value is small $\Rightarrow$ evidence to reject the null hypothesis. Conclude that quadratic model is preferable.

Extrapolation

- Be careful not to over interpret polynomial models.
- Is curvature real (based on science etc) or due to outliers.
- Polynomials are, in general, very poor for extrapolation (prediction outside the range of the data). They rapidly become large and, positive or negative.
- They may be good for prediction within the range of the x values.
- Prediction intervals get larger outside the range of x values but they still assume that the model is correct!

Alterative fitting method

- We can formulate the model using the `poly()` function.
- Makes more complicated models easier to write.

```
m2 <- lm (shot.putt ~ power.clean + I(power.clean^2),  data=Shot)    ## Original formulation
m2 <- lm (shot.putt ~ poly (power.clean, 2, raw=TRUE), data=Shot)    ## Using poly()
summary (m2)
```

```
Call:
lm(formula = shot.putt ~ poly(power.clean, 2, raw = TRUE), data = Shot)

Residuals:
     Min       1Q   Median       3Q      Max 
-1.83778 -0.61059 -0.05209  0.84770  1.82122 

Coefficients:
                                    Estimate Std. Error t value Pr(>|t|)    
(Intercept)                       -1.5318210  2.1813993  -0.702  0.48903    
poly(power.clean, 2, raw = TRUE)1  0.2827080  0.0506977   5.576 8.46e-06 ***
poly(power.clean, 2, raw = TRUE)2 -0.0010400  0.0002824  -3.682  0.00111 ** 
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 1.039 on 25 degrees of freedom
Multiple R-squared:  0.8687, Adjusted R-squared:  0.8582 
F-statistic: 82.68 on 2 and 25 DF,  p-value: 9.54e-12
```