MIE237

Neil Montgomery

2016-03-15

Multiple regression so far

- Completed:
  - Model basics, single paramter inference, confidence and prediction intervals
- Still to come:
  - Higher order terms and dummy variables
  - Model selection
  - Assumptions and plots revisited

Model selection is hard

- Model selection is a computationally intensive process, but there is no reliable algorithm. (It turns out model selection is "unstable".)
- Some plausible (and legitimate) criteria include:
  - R^2 and variations, and other single number summaries
  - Small p-values
  - Good diagnostic plots
  - Parsimony (smaller models might be better)
  - Predictive accuracy (the main criteria in "machine learning")

Higher order terms

(Note 1: This is not a textbook topic in its own right but is discussed on pp. 447 and here and there in section 12.8)

(Note 2: This topic is also being used to illustrate model selection challenges.)

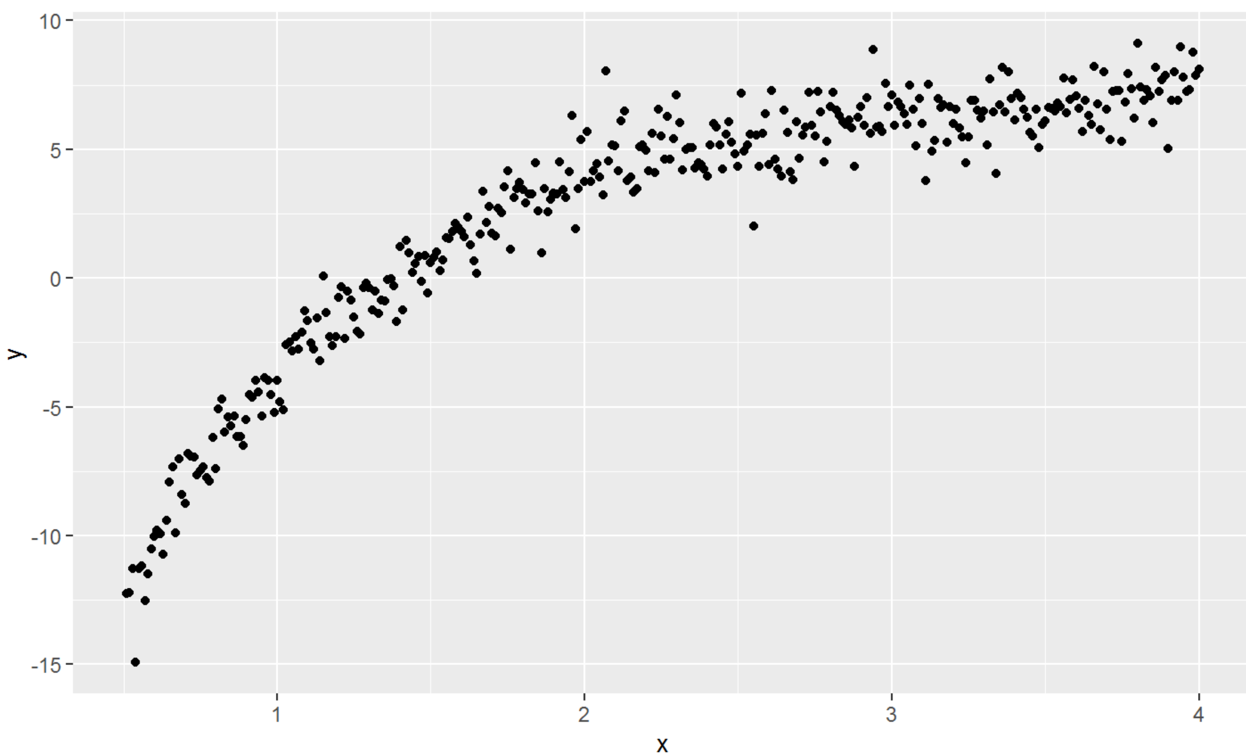
A "higher order term" in a regression model is just a product of other variables in the model. The polynomial model is an example of a model with higher order terms:

$$y_i = \beta_0 + \beta_1 x_i + \beta_2 x_i^2 + \cdots + \beta_k x_i^k + \varepsilon_i$$

Polynomial models are mainly used to fit a nonlinear relationship between a y and an x variable. To illustrate the concept I will simulate data from this model:

$$y = x^3 - 9x^2 + 28x - 24 + \varepsilon, \quad \varepsilon \sim N(0, 1)$$

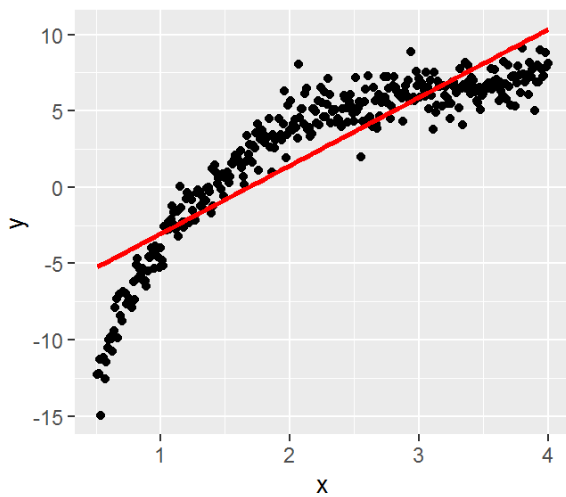
Plot of the data

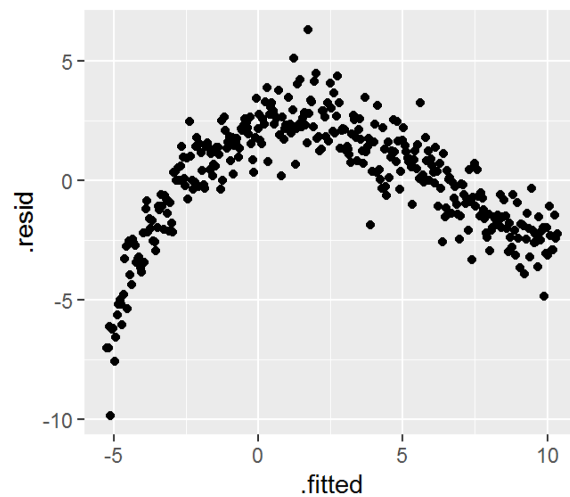


Polynomial degree 1 fit

```
##          term  estimate std.error statistic       p.value
## 1 (Intercept) -7.510257 0.3150228 -23.84036  3.705621e-75
## 2           x  4.462139 0.1274879  35.00049 4.865382e-116
```

```
##   r.squared
## 1 0.7787715
```

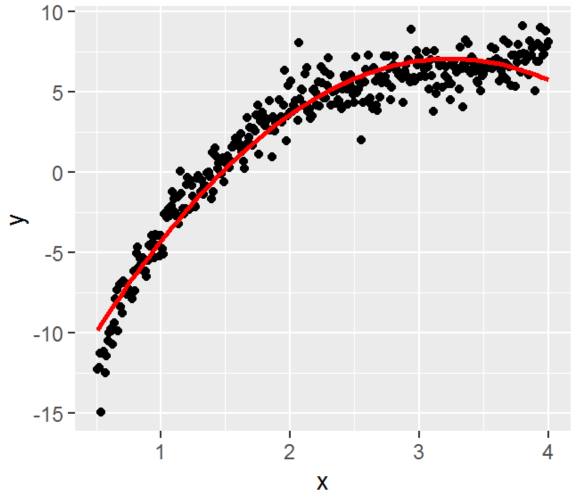


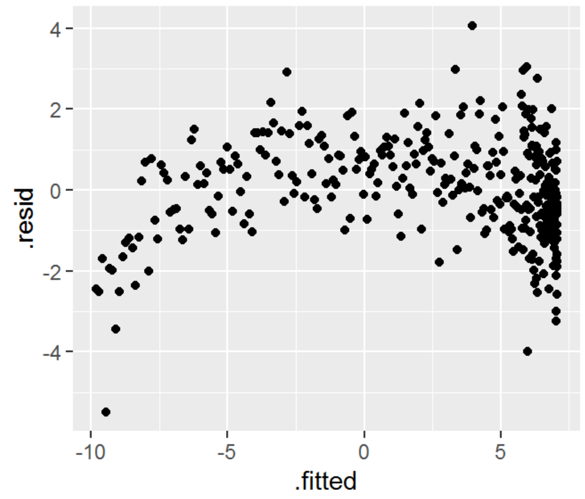


Polynomial degree 2 fit

```
##          term   estimate std.error statistic       p.value
## 1 (Intercept) -16.715803 0.3335883 -50.10908 6.051701e-161
## 2           x  14.677436 0.3308838  44.35828 4.734231e-145
## 3      I(x^2)  -2.265033 0.0719368 -31.48642 9.571907e-104
```

```
##   r.squared
## 1  0.942643
```

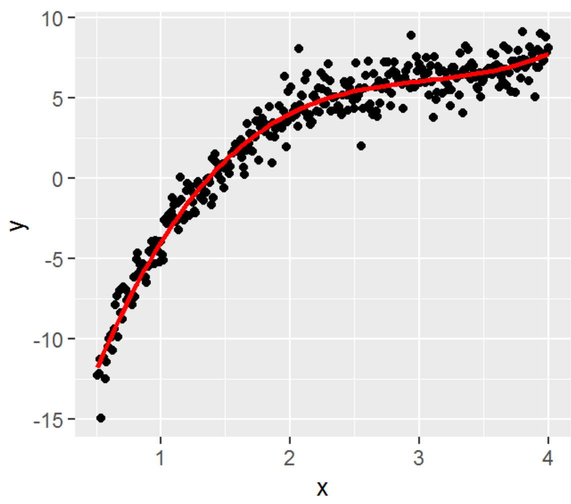


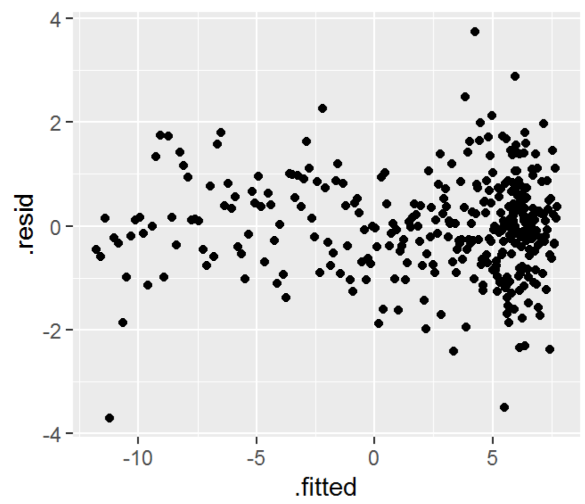


Polynomial degree 3 fit

```
##          term    estimate std.error statistic       p.value
## 1 (Intercept) -23.6044633 0.5334916 -44.24524 1.696732e-144
## 2           x  27.2987982 0.8907012  30.64866 1.258211e-100
## 3      I(x^2)  -8.6285739 0.4333223 -19.91260  2.487263e-59
## 4      I(x^3)   0.9406565 0.0635095  14.81127  8.746098e-39
```

```
##   r.squared
## 1 0.9648984
```

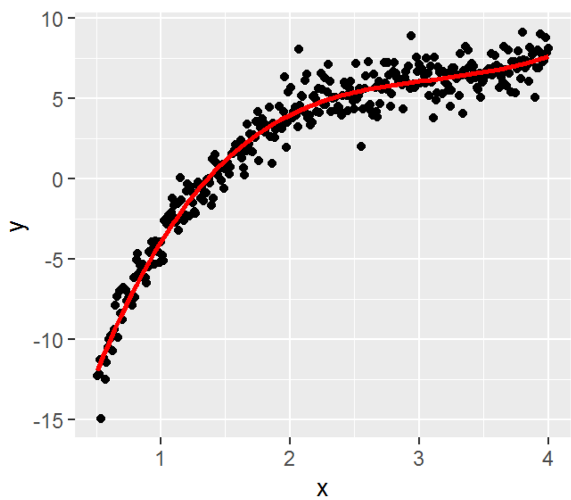


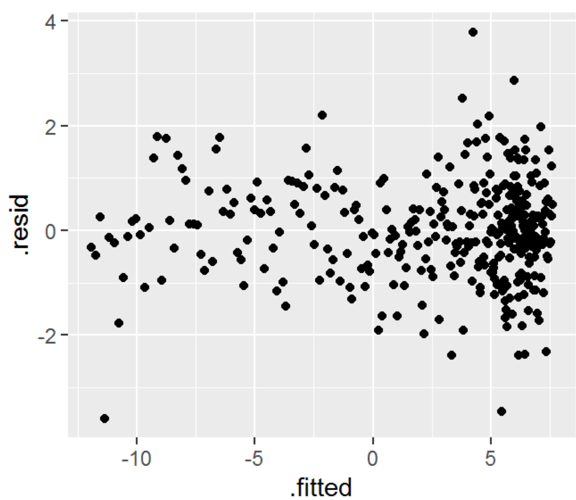


Polynomial degree 4 fit

```
##          term     estimate  std.error   statistic      p.value
## 1 (Intercept) -24.41522088 1.09769498 -22.2422634 1.280801e-68
## 2           x  29.37103545 2.60858333  11.2593817 2.947514e-25
## 3      I(x^2) -10.32669984 2.05529759  -5.0244305 8.121081e-07
## 4      I(x^3)   1.48994691 0.65296648   2.2818122 2.310898e-02
## 5      I(x^4)  -0.06089694 0.07204745  -0.8452338 3.985661e-01
```

```
##   r.squared
## 1  0.964971
```

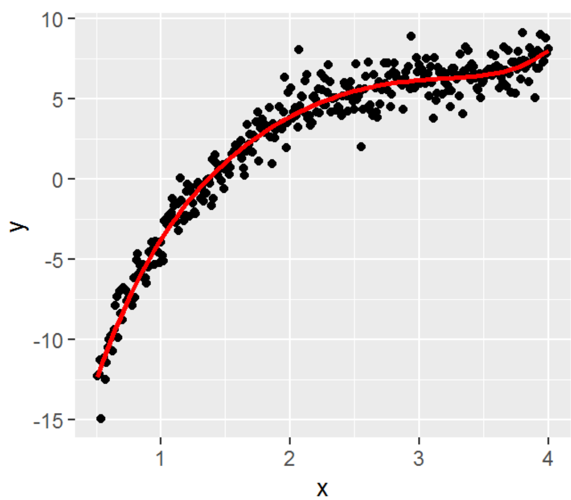


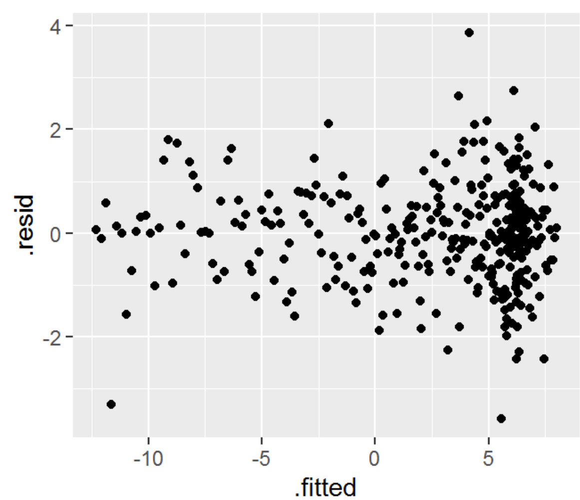


Polynomial degree 5 fit

```
##          term    estimate  std.error  statistic      p.value
## 1 (Intercept) -29.2106944 2.25962660 -12.927222 1.980081e-31
## 2           x  45.0648603 6.97635914   6.459653 3.583782e-10
## 3      I(x^2) -28.3929868 7.73111567  -3.672560 2.783343e-04
## 4      I(x^3)  10.8430733 3.91457494   2.769924 5.911236e-03
## 5      I(x^4)  -2.2834825 0.92015727  -2.481622 1.355558e-02
## 6      I(x^5)   0.1971251 0.08136334   2.422775 1.591825e-02
```

```
##   r.squared
## 1 0.9655586
```





"Overall" F tests for degrees 3 and 5

Degree 3:

source	df	sumsq	ms	F	p-value
Regression	3	8814.08	2938.03	3170.37	0.00
Error	346	320.64	0.93		

Degree 5:

source	df	sumsq	ms	F	p-value
Regression	5	8820.11	1764.02	1928.80	0.00
Error	344	314.61	0.91		

Polynomial example comments

As expected, the 3rd degree polynomial model is the best model.

Note that 4th and beyond are still perfectly good predictive models!(Despite some "individual" p-values being large…)

Always remember the correct interpretation of these p-values.

"Overall" F test can show strong evidence of a model even with "individual" p-values small.

These apparent issues are caused (in this case) by powers of x being highly correlated over the range of the data.

the sample correlation coeffients

Here is a matrix of sample correlation coefficients among the first five powers of x over its range $[0.51, 4]$.

1.00000	0.98051	0.94067	0.89676	0.85478
0.98051	1.00000	0.98844	0.96377	0.93468
0.94067	0.98844	1.00000	0.99282	0.97666
0.89676	0.96377	0.99282	1.00000	0.99522
0.85478	0.93468	0.97666	0.99522	1.00000