MIE237

Neil Montgomery

2016-03-18

midterm 2 information

Details

2016-03-22 from 15:10 to 16:00 in rooms EX310 and EX320.

Rooms divided same as last time.

Topics: simple and multiple regression (not including model selection)

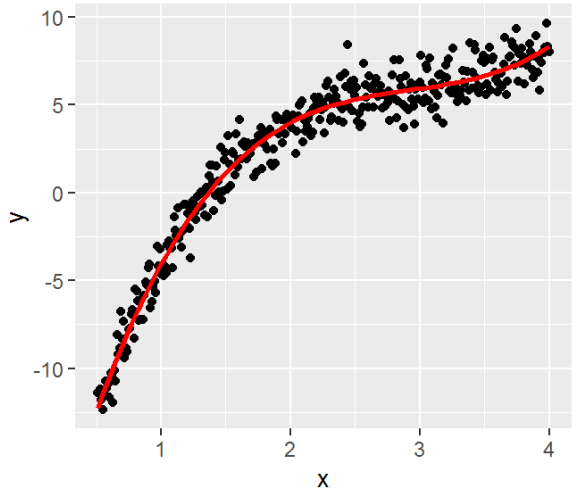
No lab next week. Hootan will hold office hours on Tuesday in lieu of a formal tutorial.

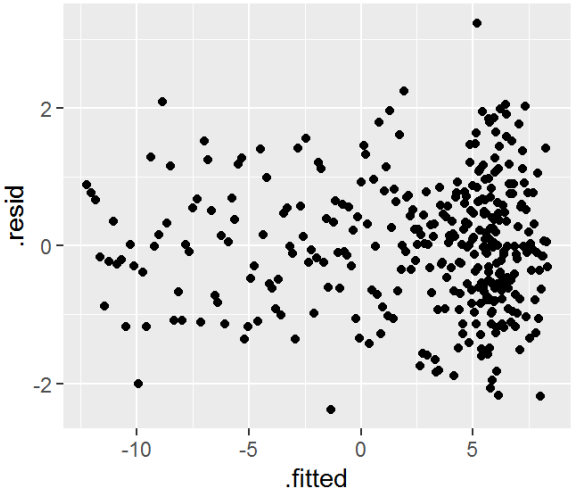
I will have a Q&A free-for-all and past test question review in the latter part of today's class.

Polynomial degree 3 fit

```
##          term   estimate  std.error statistic       p.value
## 1 (Intercept) -24.830708 0.51932067 -47.81383 1.616539e-154
## 2           x  29.294008 0.86704181  33.78615 1.300464e-111
## 3      I(x^2)  -9.636614 0.42181208 -22.84575  4.407080e-71
## 4      I(x^3)   1.096436 0.06182252  17.73522  1.620614e-50
```

```
##   r.squared
## 1 0.9679878
```

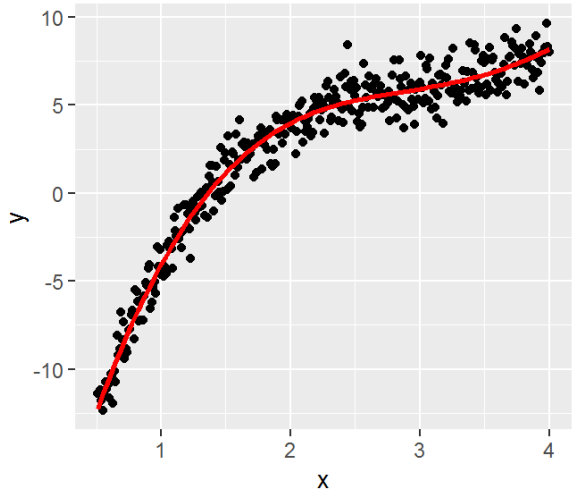


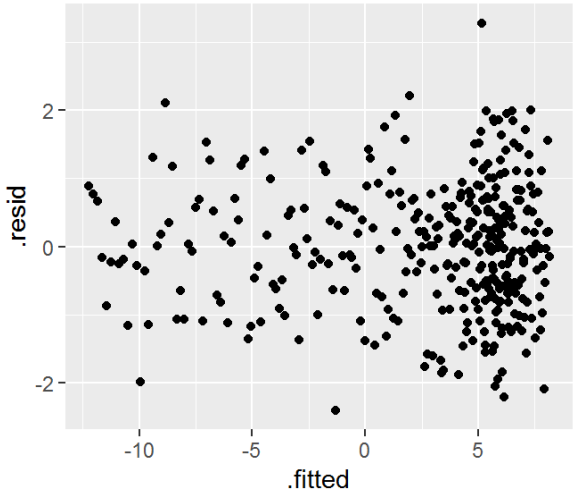


Polynomial degree 5 fit

```
##          term     estimate  std.error   statistic      p.value
## 1 (Intercept) -24.38645282 2.21865792 -10.9915335 2.744626e-24
## 2           x  27.44953214 6.84987267   4.0073055 7.532159e-05
## 3      I(x^2)  -7.05126088 7.59094491  -0.9289042 3.535901e-01
## 4      I(x^3)  -0.46375212 3.84360084  -0.1206556 9.040342e-01
## 5      I(x^4)   0.41755305 0.90347415   0.4621638 6.442560e-01
## 6      I(x^5)  -0.04066406 0.07988817  -0.5090123 6.110699e-01
```

```
##   r.squared
## 1 0.9680434
```





"Overall" F tests for degrees 3 and 5

Degree 3:

source	df	sumsq	ms	F	p-value
Regression	3	9187.40	3062.47	3487.46	0.00
Error	346	303.84	0.88		

Degree 5:

source	df	sumsq	ms	F	p-value
Regression	5	9187.93	1837.59	2084.12	0.00
Error	344	303.31	0.88		

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

Another higher order term: interaction

"How long does it take to install a boiler?"

The time is plausibly related to both the size of the boiler and its "rating" (pressure), all else being equal.

Consider the following table:

		Size	
		Low	High
kPa	Low	20h	30h
	High	30h	

What would it mean for the missing entry to be: 30, 40, or 50?

Interaction

If the effect of x_1 on y depends on the value of x_2, wesay there is an between x_1 and x_2.

Model:

$$y_i = \beta_0 + \beta_1 x_{1i} + \beta_2 x_{2i} + \beta_3 x_{1i} x_{2i} + \varepsilon_i$$

When x_1 increases by one unit, y increases by $\beta_1 + \beta_3 x_2$ units.

When x_2 increases by one unit, y increases by $\beta_2 + \beta_3 x_1$ units.

Boiler example

```
##
## Call:
## lm(formula = hours ~ cap + kPa + cap * kPa, data = .)
##
## Residuals:
##      Min       1Q   Median       3Q      Max
## -2964.74  -914.17    53.46  1028.66  2839.99
##
## Coefficients:
##               Estimate Std. Error t value Pr(>|t|)
## (Intercept)  3.013e+03  7.324e+02   4.114 0.000254 ***
## cap          3.786e+00  2.135e+00   1.774 0.085659 .
## kPa         -2.218e-01  1.571e-01  -1.412 0.167689
## cap:kPa      4.988e-04  2.234e-04   2.233 0.032662 *
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 1472 on 32 degrees of freedom
## Multiple R-squared:  0.7287, Adjusted R-squared:  0.7033
## F-statistic: 28.65 on 3 and 32 DF,  p-value: 3.43e-09
```

Boiler (without the interaction term)

```
##
## Call:
## lm(formula = hours ~ cap + kPa, data = .)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -3498.9  -822.2   132.5  1142.8  4287.7
##
## Coefficients:
##             Estimate Std. Error t value Pr(>|t|)
## (Intercept) 1.671e+03  4.430e+02   3.771 0.000642 ***
## cap         7.451e+00  1.446e+00   5.152 1.18e-05 ***
## kPa         4.782e-02  1.064e-01   0.449 0.656138
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 1559 on 33 degrees of freedom
## Multiple R-squared:  0.6864, Adjusted R-squared:  0.6674
## F-statistic: 36.12 on 2 and 33 DF,  p-value: 4.891e-09
```

Dummy variables

Categorical ("factor") variables can be included in regression models. But they must somehow be encoded as numbers. The way to do this is to use "dummy variables" that take on the values 0 and 1.

If there are I "levels" of the factor variable, you need $I - 1$ dummy variables for the encoding. Then include them in the regression model as usual.